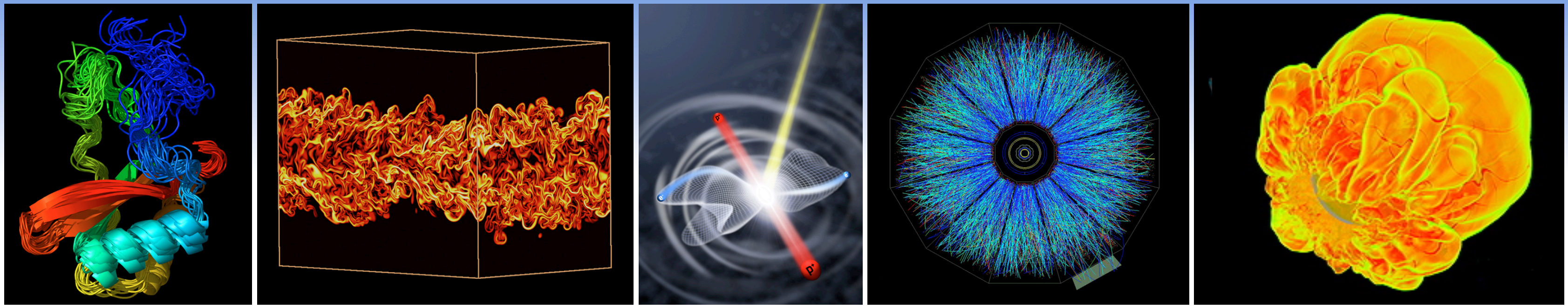


Extreme Scale Computing and Networking Environment

John Shalf

National Energy Research Supercomputing Center
Lawrence Berkeley National Laboratory

Terabit Networks for Extreme Scale Science

February 16, 2011

Outline

- **Technology Challenges for Next Decade**
  - **Challenges:** Power, logic, and cost of data movement
  - **Opportunities**: silicon photonics and SoC integration

- **Some Applications Drivers for High Performance Networking**
  - **Challenges**: UQ for Predictive Modeling, support for large experiments, data reanalysis
  - **Opportunities**: Data intensive computing for UQ, data assimilation, and shot planning for large experiments

A Few Words about the Exascale Computing Platforms

- Two “*associations*” of labs to direct development of exascale systems
  - Cooperation between NNSA and SC

- Each *association* puts out RFP for “vendor partners”
  - public/private partnership for platform development

- Two platform deliveries per association
  - 2 systems per delivery: one NNSA and one for SC
  - 2015: 0.3 Exaflops @ 15MW
  - 2018: 1 Exaflop @ 20MW
  - That’s a total of 8 systems

Traditional Sources of Performance Improvement are Flat-Lining

- **Moore's Law is alive and well**
- **15 years of *exponential* clock speed growth has ended**
- **How to use the transistors?**
  - Industry Response: #cores per chip doubles every 18 months *instead* of clock frequency!
  - *Technology disruption will force redesign of many aspects of our computing environment*

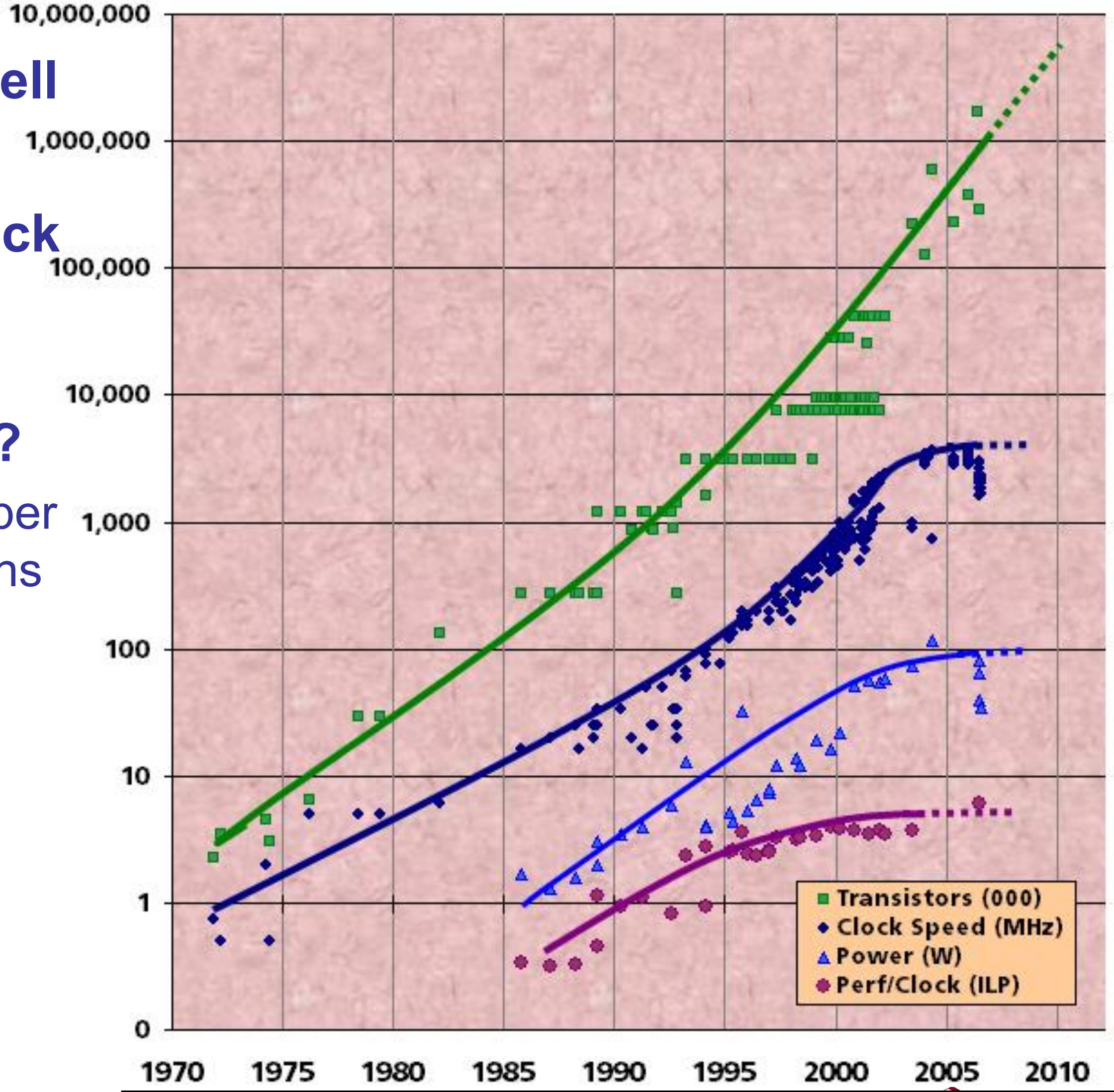


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Technology Disruptions on the Path to Exascale

- **Gigaflops to Teraflops was highly disruptive**
  - Moved from vector machines to MPPs with message passing
  - Required new algorithms and software

- **Teraflops to Petaflops was *not* very disruptive**
  - Continued with MPI+Fortran/C/C++ with incremental advances

- **Petaflops to Exaflops will be highly disruptive**
  - No clock increases → hundreds of simple "cores" per chip
  - Less memory and bandwidth → cores are not MPI engines
  - x86 too energy intensive → more technology diversity (GPUs/ accel.)
  - Programmer controlled memory hierarchies likely

- **Computing at every scale will be *transformed (not just exascale)***

Systems	2009	2015 +1/-0	2018 +1/-0
System peak	**2 Peta**	**100-300 Peta**	**1 Exa**
Power	**6 MW**	**~15 MW**	**~20 MW**
System memory	0.3 PB	5 PB	64 PB (+)
Node performance	125 GF	0.5 TF or 7 TF	2 TF or 10TF
Node memory BW	25 GB/s	0.2TB/s or 0.5TB/s	0.4TB/s or 1TB/s
Node concurrency	12	O(100)	O(1k) or 10k
Total Node Interconnect BW	3.5 GB/s	100-200 GB/s 10:1 vs memory bandwidth 2:1 alternative	200-400GB/s (1:4 or 1:8 from memory BW)
System size (nodes)	18,700	50,000 or 500,000	O(100,000) or O(1M)
Total concurrency	225,000	O(100,000,000) *O(10)-O(50) to hide latency	O(billion) * O(10) to O (100) for latency hiding
Storage	15 PB	150 PB	500-1000 PB (>10x system memory is min)
IO	0.2 TB	10 TB/s	60 TB/s (how long to drain the machine)
MTTI	days	O(1day)	O(1 day)

The REAL Exascale Constraints

First Generation

- 300PF
- 15MW
- $200M
- Deliver by 2015

Second Generation

- 1 Exaflop
- 20MW
- $200M
- Deliver by 2018

Do not get caught up in the tyranny of the spreadsheet!
all parameters are movable (with consequences)
co-design: optimize movable parameters

Changing Notion of “System Balance”

- If you pay 5% more to double the FPUs and get 10% improvement, it’s a win (despite lowering your % of peak performance)

- If you pay 2x more on memory BW (power or cost) and get 35% more performance, then it’s a net loss (even though % peak looks better)

- *Real example: we can give up ALL of the flops to improve memory bandwidth by 20% on the 2018 system*

- We have a fixed budget (power and $s)
  - Sustained to peak FLOP rate is *wrong* metric if FLOPs are cheap
  - Balance involves balancing your checkbook & balancing your power budget
  - Requires a application co-design make the right trade-offs

The Challenge

Where do we get a 1000x improvement in performance with only a 10x increase in power?

How do you achieve this in 10 years with a finite development budget?

Loss-Leaders: Transistors and Wires

CMOS Logic and Cost of Moving Data

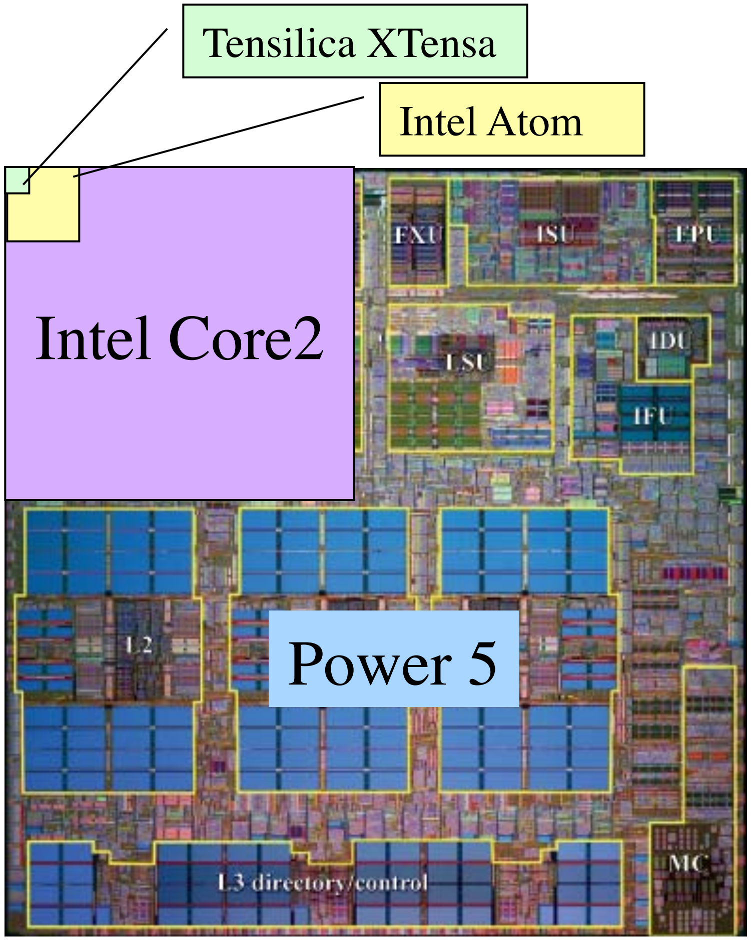
Processors: What are the problems?

(Lessons from the Berkeley View)

- **Current Hardware/Lithography Constraints**
  - **Power limits leading edge chip designs**
    - Intel Tejas Pentium 4 cancelled due to power issues
  - **Yield on leading edge processes dropping dramatically**
    - IBM quotes yields of 10 – 20% on 8-processor Cell
  - **Design/validation leading edge chip is becoming unmanageable**
    - Verification teams > design teams on leading edge processors

- **Solution: Small Is Beautiful**
  - **Simpler (5- to 9-stage pipelined) CPU cores**
    - Small cores not much slower than large cores
  - **Parallel is energy efficient path to performance:CV^2F**
    - Lower threshold and supply voltages lowers energy per op
  - **Redundant processors can improve chip yield**
    - Cisco Metro 188 CPUs + 4 spares; Sun Niagara sells 6 or 8 CPUs
  - **Small, regular processing elements easier to verify**

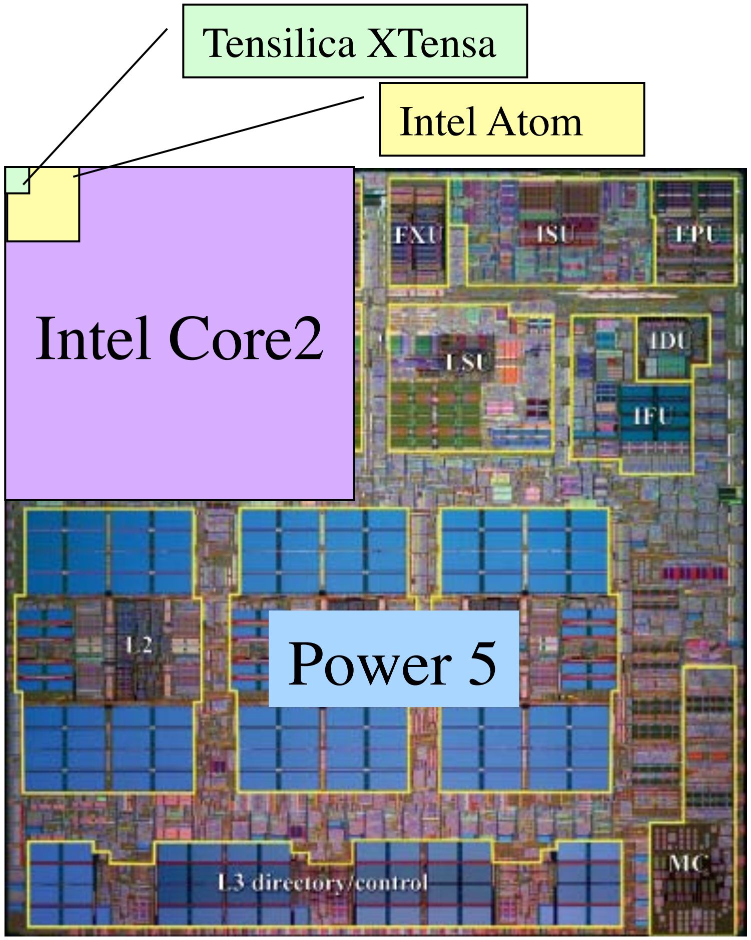
Low-Power Design Principles



- Cubic power improvement with lower clock rate due to V^2F
- Slower clock rates enable use of simpler cores
- Simpler cores use less area (lower leakage) and reduce cost
- Tailor design to application to REDUCE WASTE

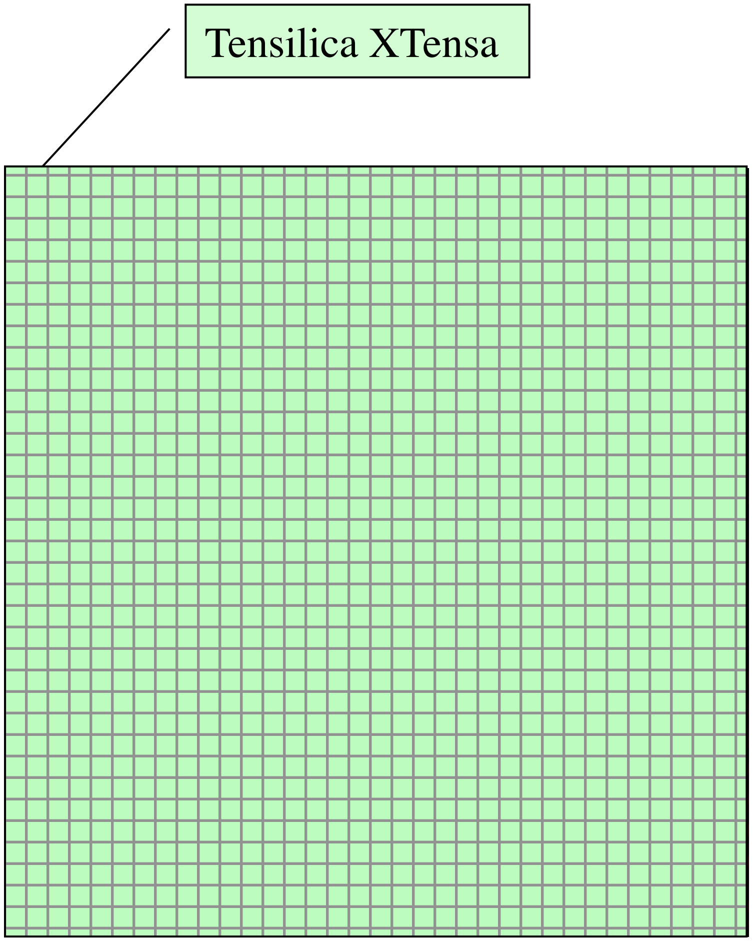
This is how iPhones and MP3 players are designed to maximize battery life and minimize cost

Low-Power Design Principles



- Power5 (server)
  - 120W@1900MHz
  - Baseline
- Intel Core2 sc (laptop) :
  - 15W@1000MHz
  - *4x more FLOPs/watt than baseline*
- Intel Atom (handhelds)
  - 0.625W@800MHz
  - 80x more
- Tensilica XTensa DP (Moto Razor) :
  - 0.09W@600MHz
  - 400x more *(80x-120x sustained)*

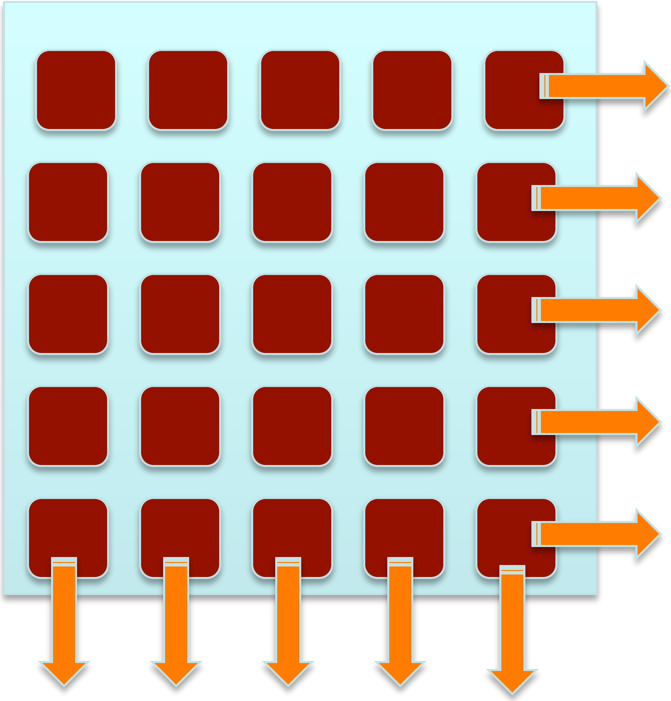
Low Power Design Principles



- Power5 (server)
  - 120W@1900MHz
  - Baseline
- Intel Core2 sc (laptop) :
  - 15W@1000MHz
  - *4x more FLOPs/watt than baseline*
- Intel Atom (handhelds)
  - 0.625W@800MHz
  - 80x more
- Tensilica XTensa DP (Moto Razor) :
  - 0.09W@600MHz
  - 400x more (80x-100x sustained)

Even if each simple core is 1/4th as computationally efficient as complex core, you can fit hundreds of them on a single chip and still be 100x more power efficient.

Future of On-Chip Architecture
(San Diego Meeting)

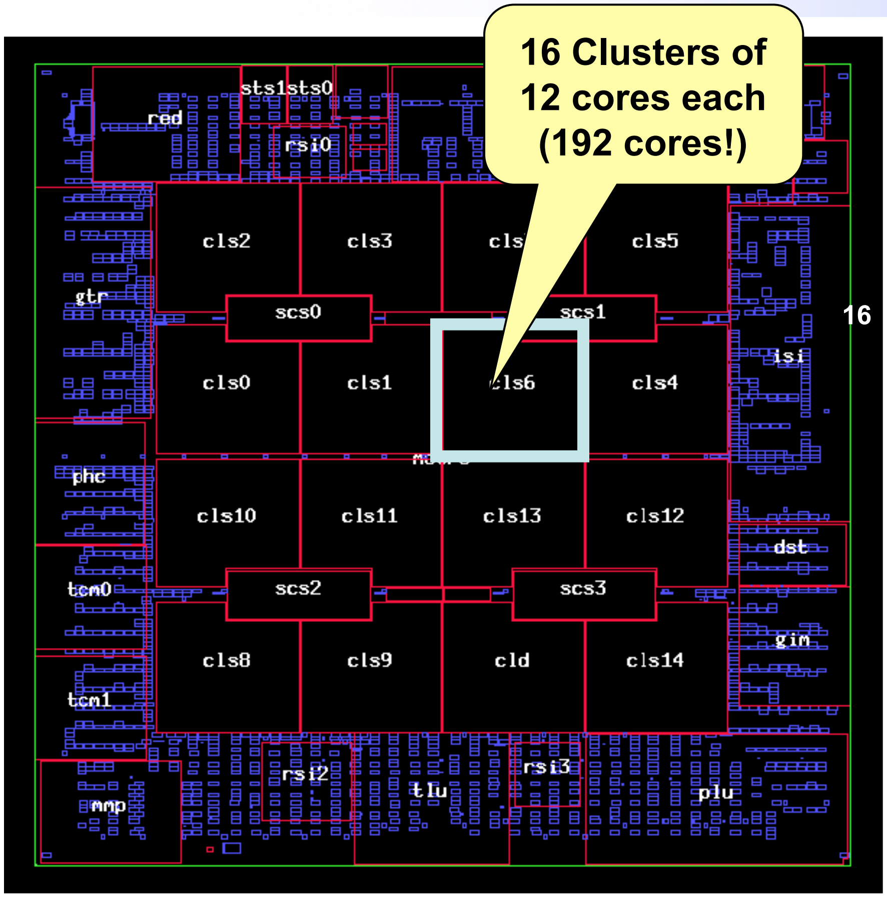


Scale-out for Planar geometry

- **~1000-10k simple cores /Chip**
  - 4-8 wide SIMD or VLIW bundles
  - Either 4 or 50+ HW threads
- **On-chip communication Fabric**
  - Low-degree topology for on-chip communication (torus or mesh)
  - *Scale cache coherence?*
  - Global (nonCC memory)
  - Shared register file (clusters)
- **Off-chip communication fabric**
  - Integrated directly on an SoC
  - Reduced component counts
  - Coherent with TLB (no pinning)

Parallel Computing Everywhere

Cisco CRS-1 Terabit Router



16 Clusters of 12 cores each (192 cores!)

Cisco Systems

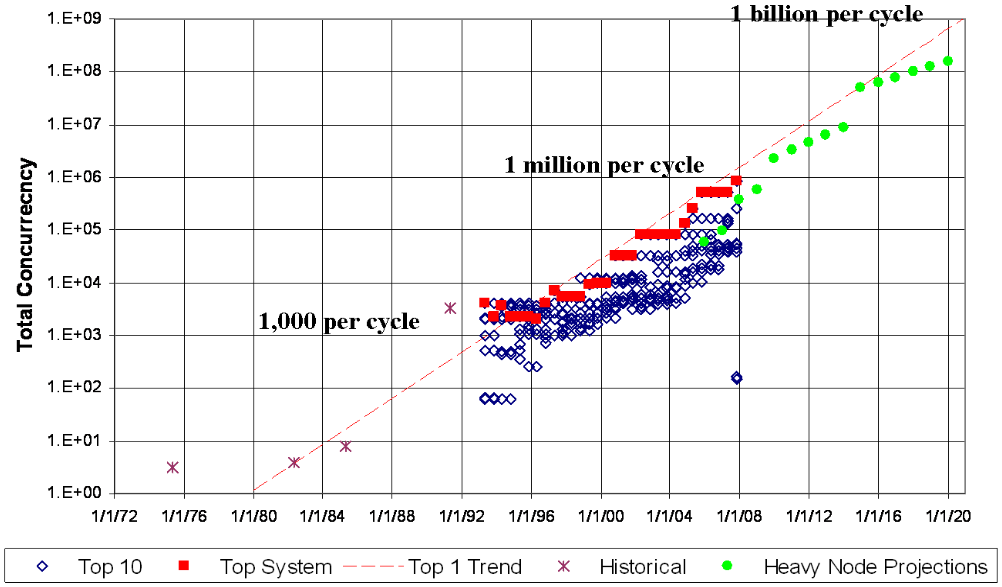
- 188+4 Xtensa general purpose processor cores per Silicon Packet Processor
- Up to 400,000 processors per system

(this is not just about HPC!!!)

Mitigates fact that we can design more logic than we can verify

Conclusion: Solving Logic Power Drives Move to Massive Parallelism

- **Future HPC must move to simpler power-efficient core designs**
  - Embedded/consumer electronics technology is central to the future of HPC
  - Convergence inevitable because it optimizes both cost and power efficiency



How much parallelism must be handled by the program?

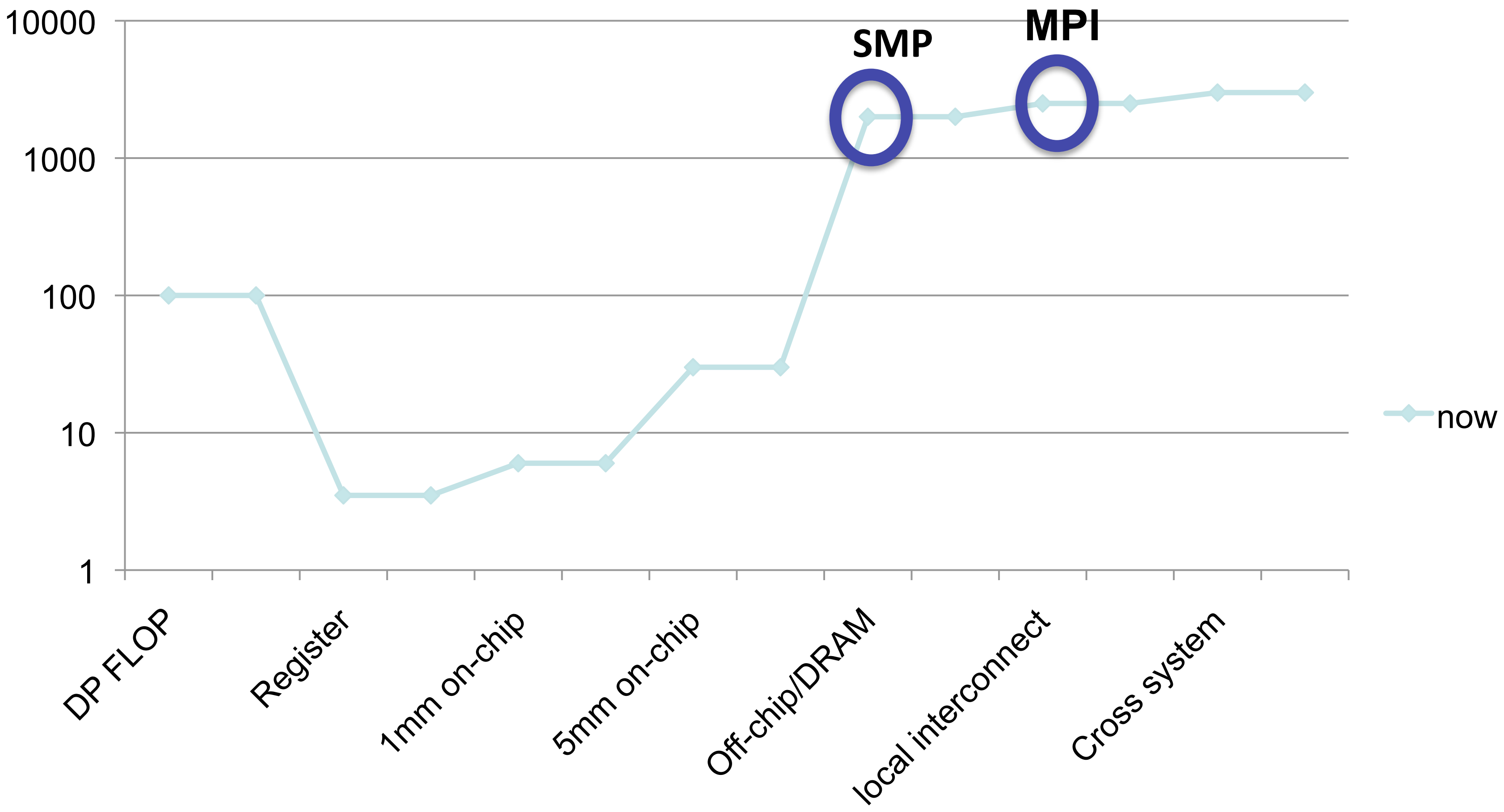
From Peter Kogge (on behalf of Exascale Working Group), “Architectural *Challenges* at the Exascale Frontier”, June 20, 2008

- **Consequence is massive on-chip parallelism**
  - A thousand cores on a chip by 2018
  - 1 Million to 1 Billion-way System Level Parallelism
  - *Must express massive parallelism in algorithms and pmodels*
  - *Must manage massive parallelism in system software*

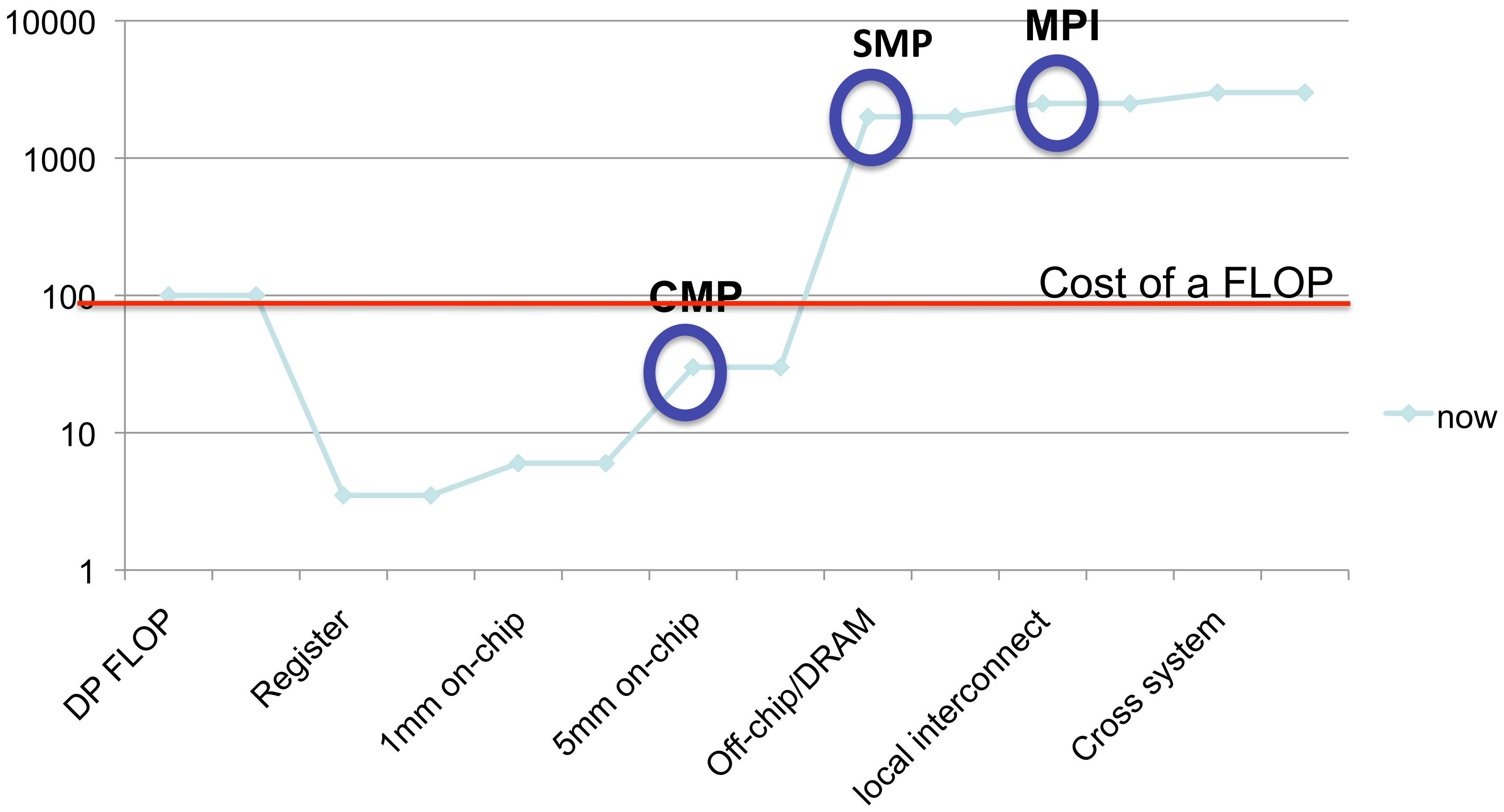
The cost of moving data

integrated optics and lambda switching

The Cost of Data Movement



The Cost of Data Movement

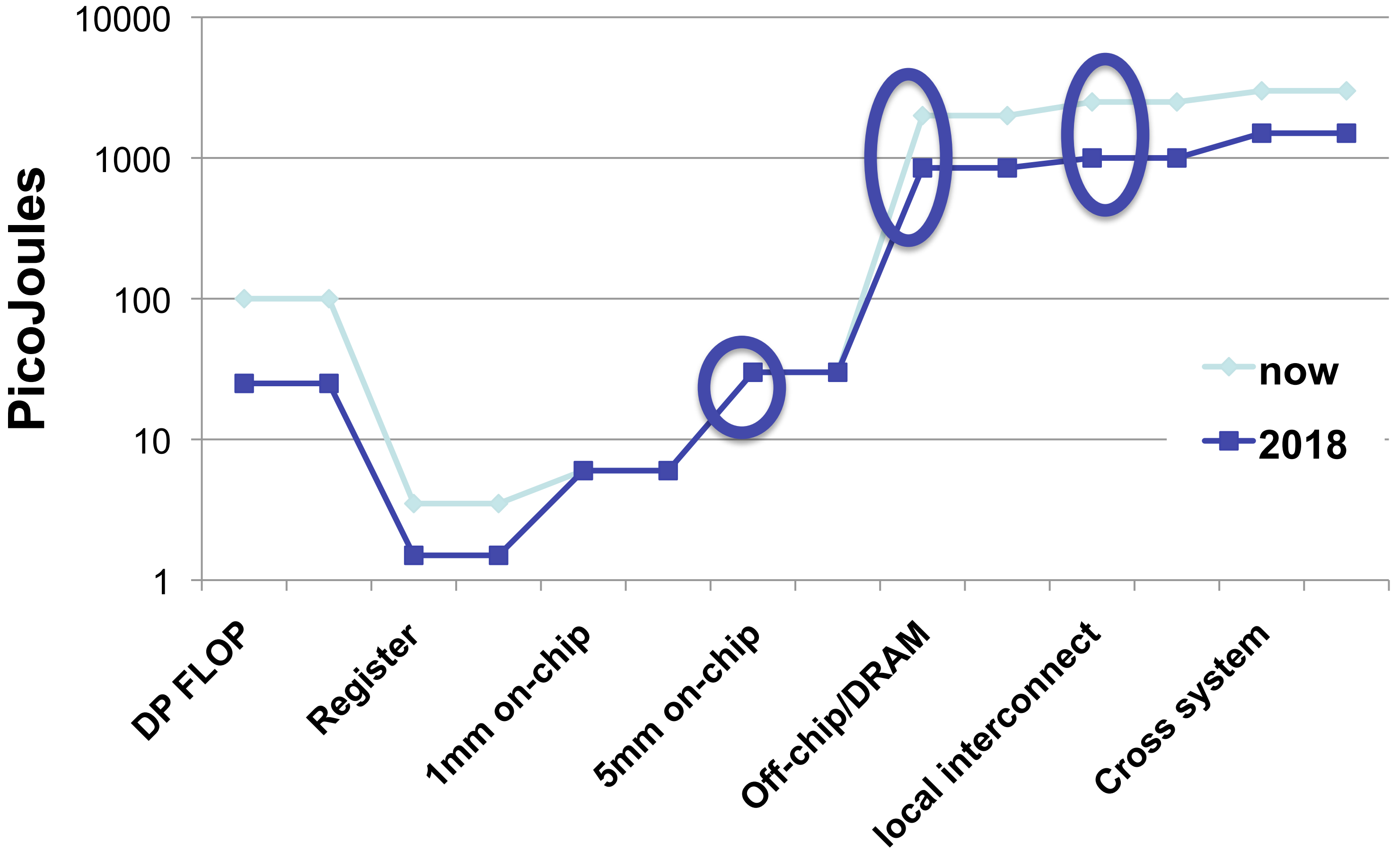


U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

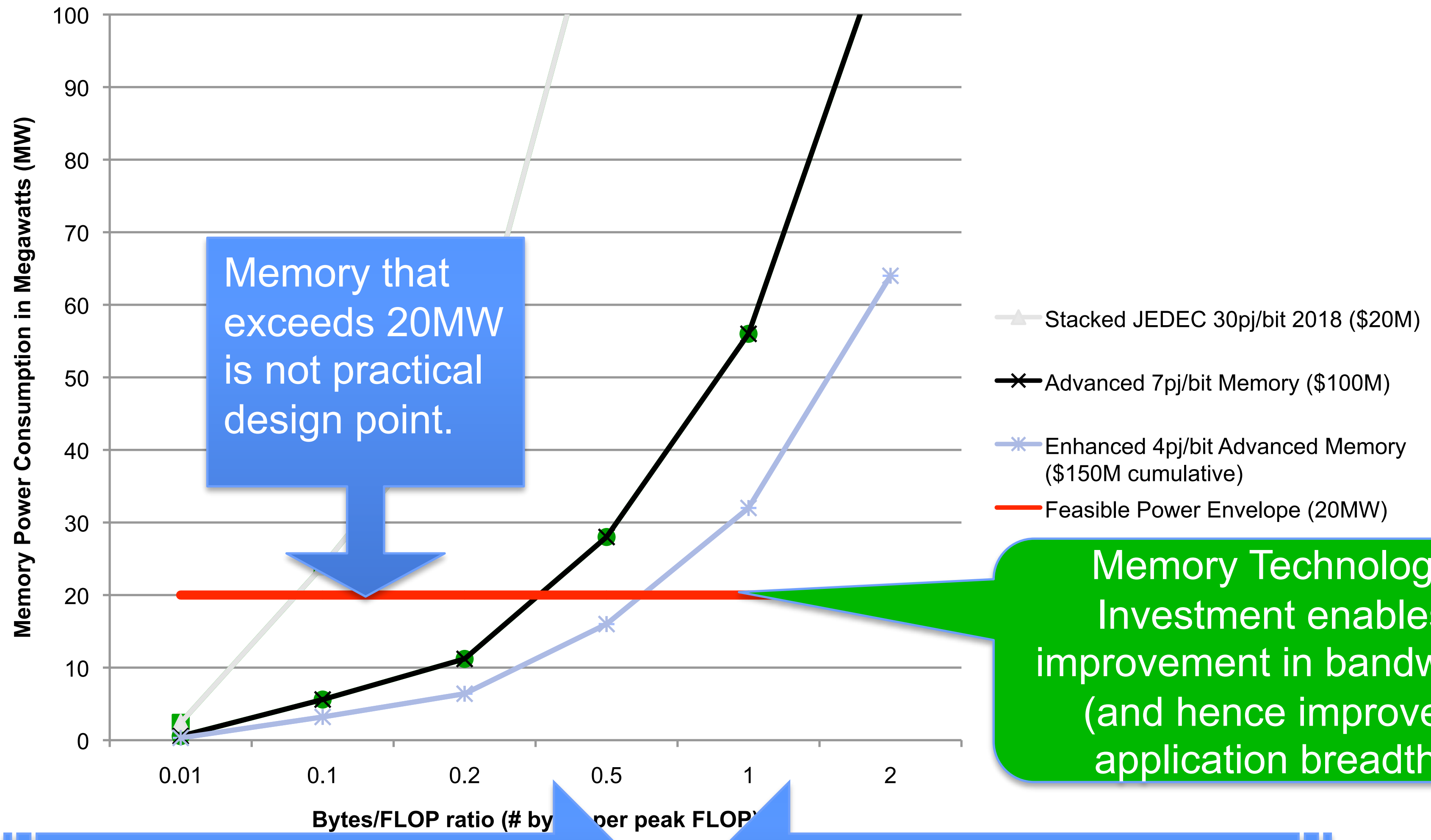
The situation will not improve in 2018

Energy Efficiency will require careful management of data locality



Important to know when you are on-chip and when data is off-chip!

Limiting Memory Bandwidth Limits System Scope



Memory that exceeds 20MW is not practical design point.

Memory Technology Investment enables improvement in bandwidth (and hence improves application breadth)

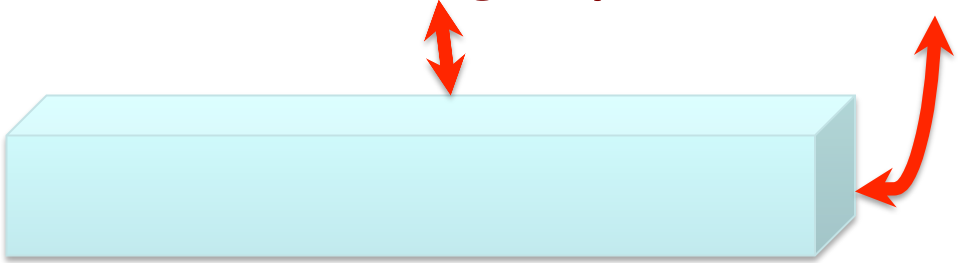
Application performance and breadth pushes us to higher BW

Power pushes us to lower bandwidth

The problem with Wires:

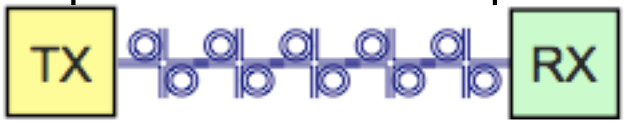
Energy to move data proportional to distance

- Cost to move a bit on copper wire:
  - **Power = bitrate * Length^2 / cross-section area**

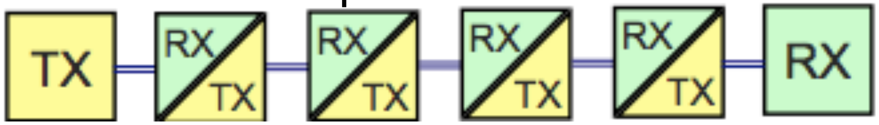


- Wire data capacity constant as feature size shrinks
- *Cost to move bit proportional to distance*
- *~1TByte/sec max feasible off-chip BW (10GHz/pin)*
- *Photonics reduces distance-dependence of bandwidth*

Photonics requires no redrive
and passive switch little power

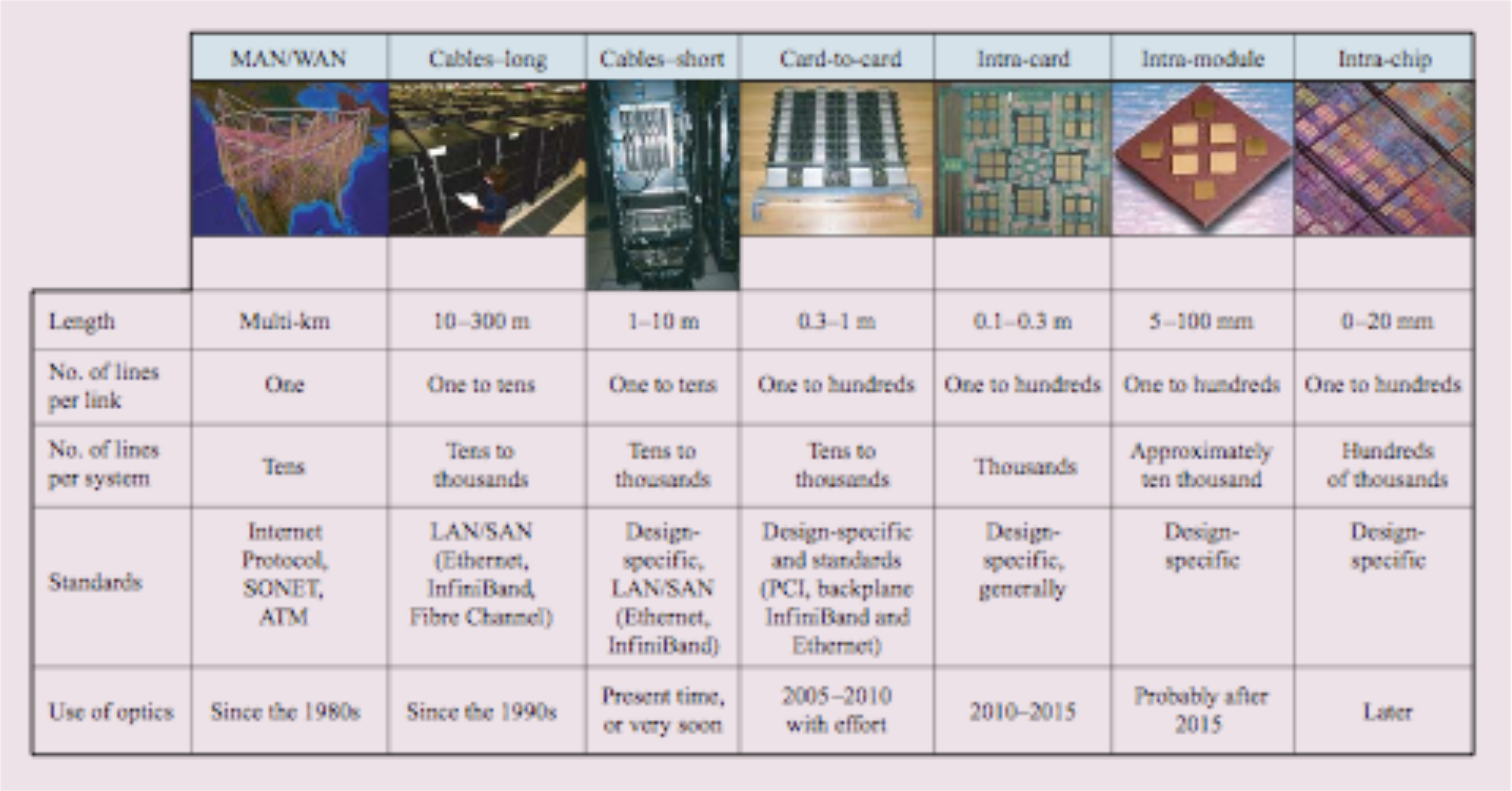


Copper requires to signal amplification
even for on-chip connections



Kash & Benner (2005)

progression towards on-chip optics



	MAN/WAN	Cables–long	Cables–short	Card-to-card	Intra-card	Intra-module	Intra-chip
Length	Multi-km	10–300 m	1–10 m	0.3–1 m	0.1–0.3 m	5–100 mm	0–20 mm
No. of lines per link	One	One to tens	One to tens	One to hundreds	One to hundreds	One to hundreds	One to hundreds
No. of lines per system	Tens	Tens to thousands	Tens to thousands	Tens to thousands	Thousands	Approximately ten thousand	Hundreds of thousands
Standards	Internet Protocol, SONET, ATM	LAN/SAN (Ethernet, InfiniBand, Fibre Channel)	Design-specific, LAN/SAN (Ethernet, InfiniBand)	Design-specific and standards (PCI, backplane InfiniBand and Ethernet)	Design-specific, generally	Design-specific	Design-specific
Use of optics	Since the 1980s	Since the 1990s	Present time, or very soon	2005–2010 with effort	2010–2015	Probably after 2015	Later

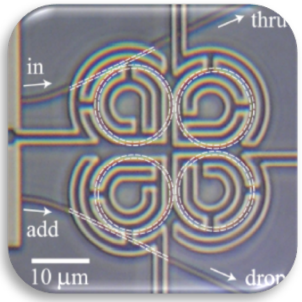
Silicon Photonics

Silicon-on-insulator (SOI) platform produces valuable photonic building blocks

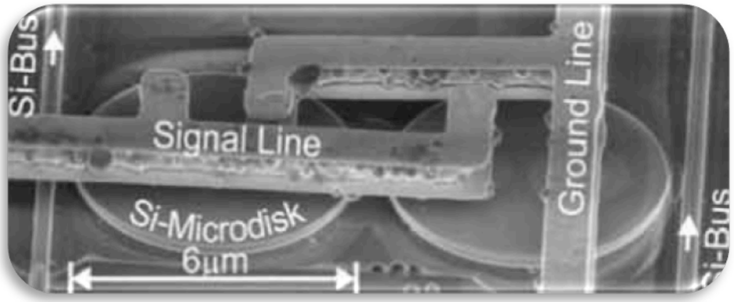
High index contrast enables high confinement, low-loss propagation, virtually lossless bending

CMOS compatibility allows monolithic integration with advanced microelectronics

Many active and passive functionalities have already been demonstrated

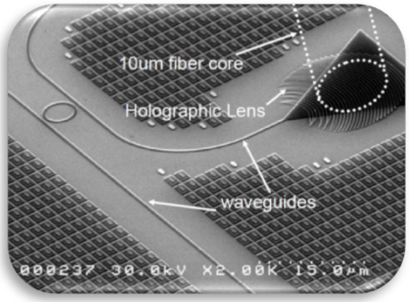


MIT



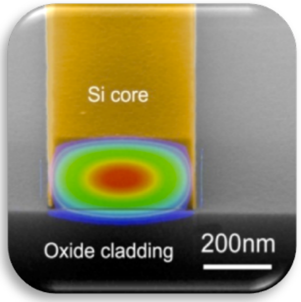
Sandia

Ghent

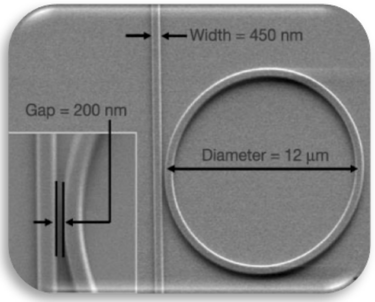


Luxtera

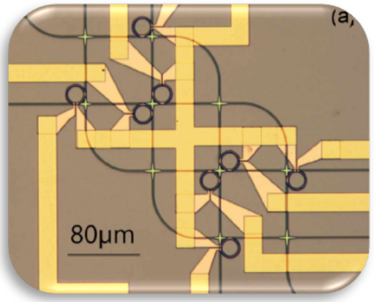
IBM



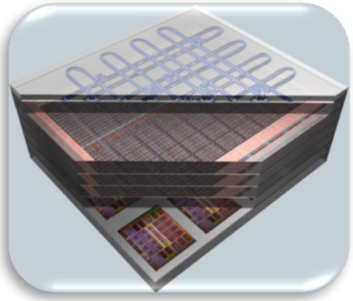
IBM



Cornell

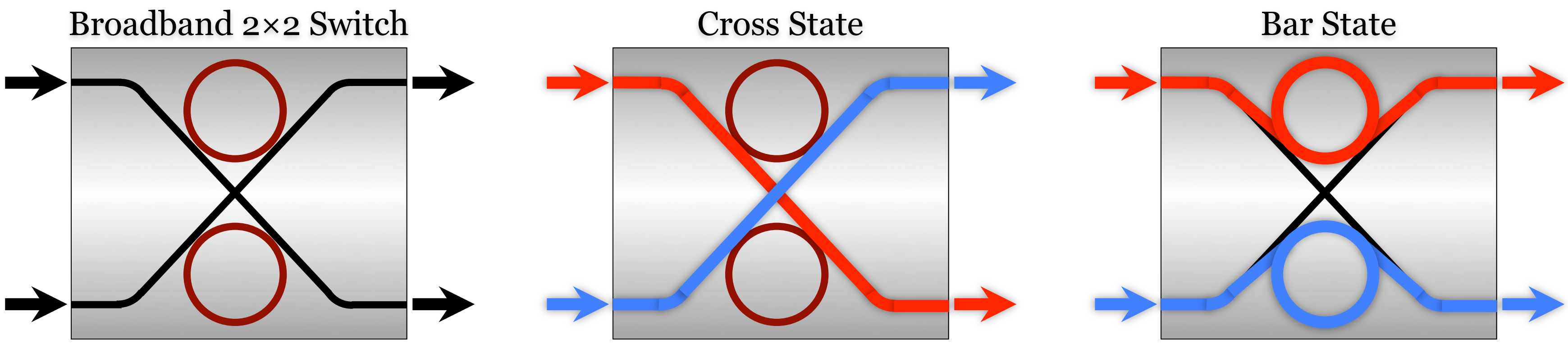


Cornell/Columbia

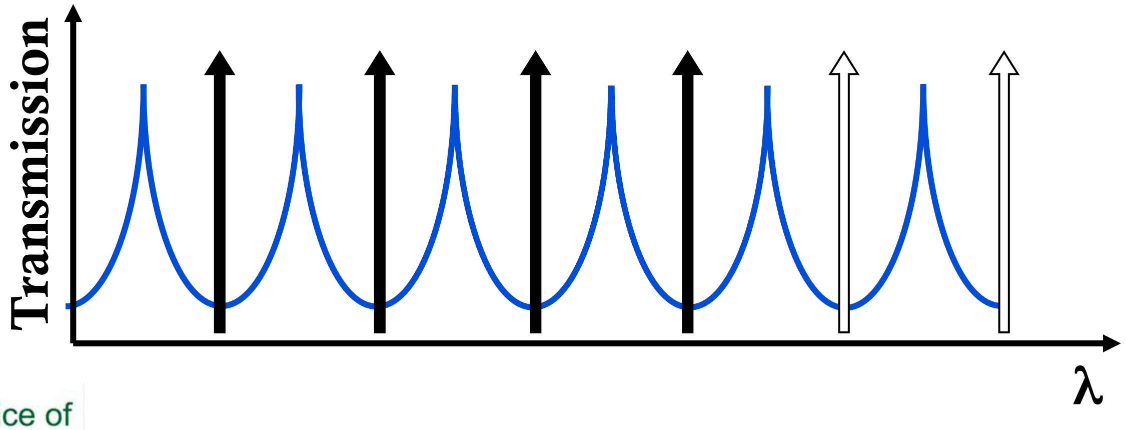


Columbia

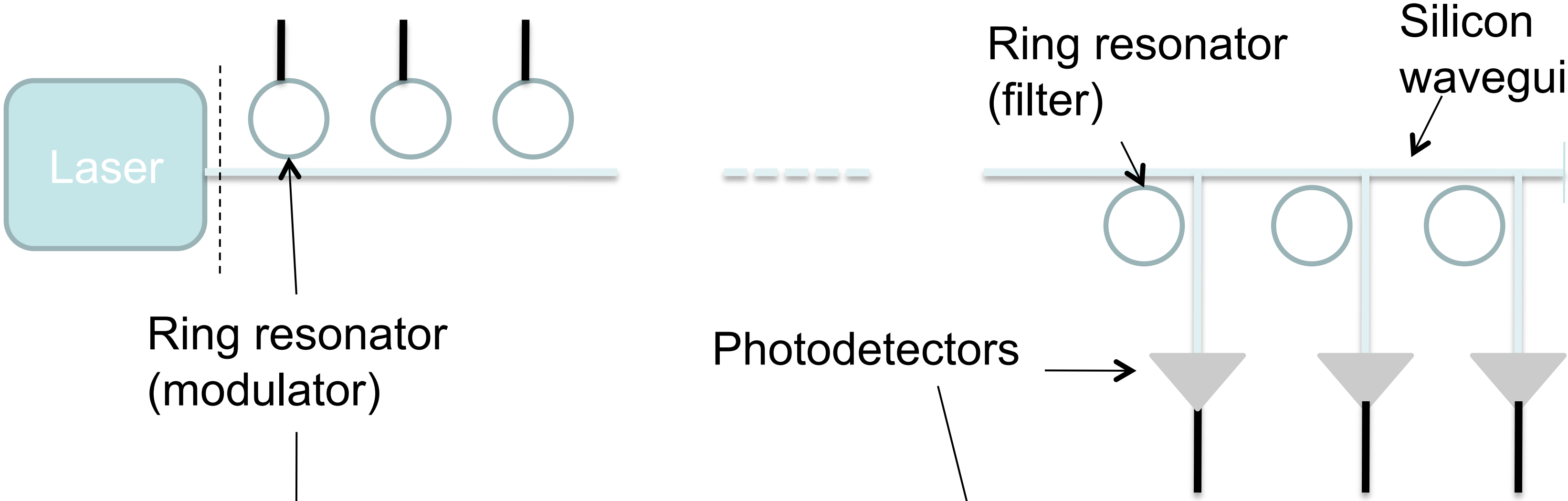
Switching Building Blocks



B. G. Lee, ECOC 2008



NanoPhotonic Devices



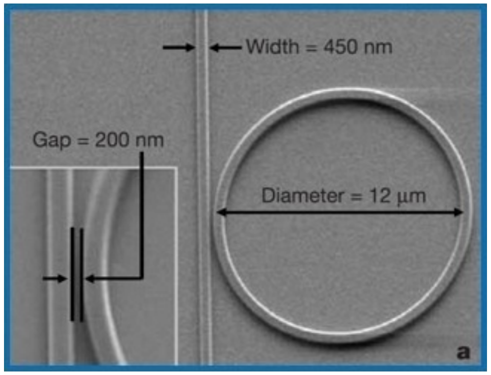
Ring resonator (filter)

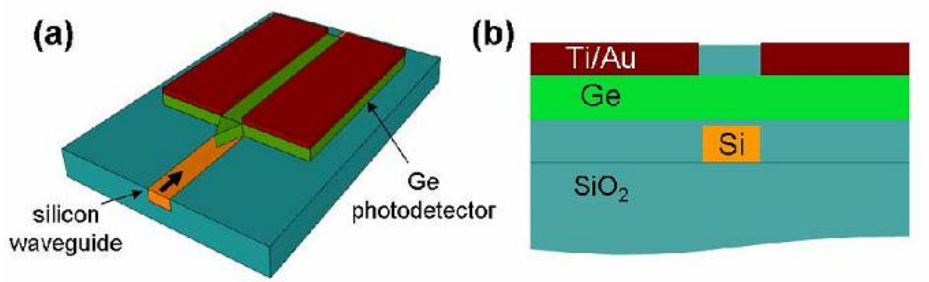
Silicon waveguides

Ring resonator (modulator)

Photodetectors

Electronic data

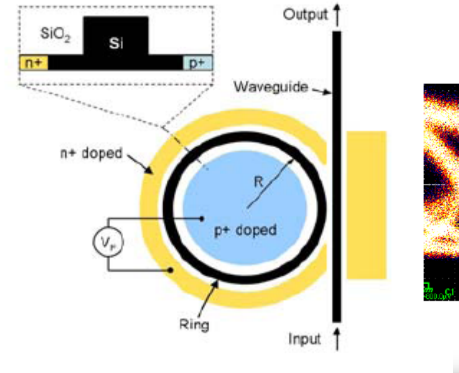




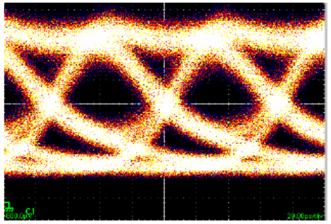
L. Chen, *OE*, 2008

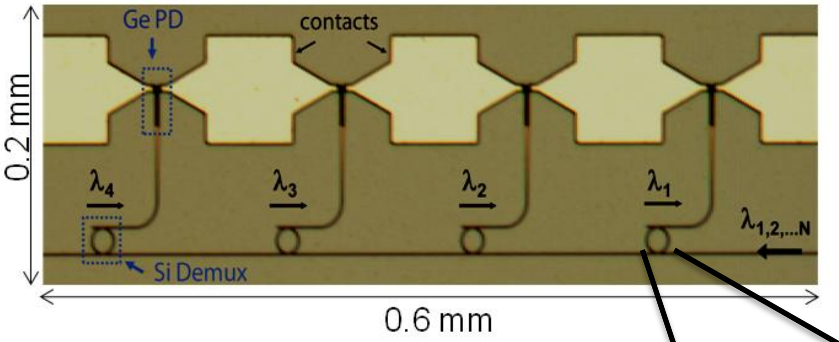
Energy Efficient E/O: Silicon Photonic WDM Data Modulation and Reception

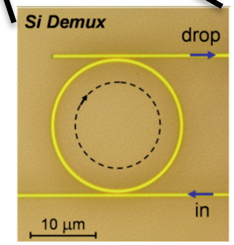
Lipson, Nanophotonics Group
Cornell University



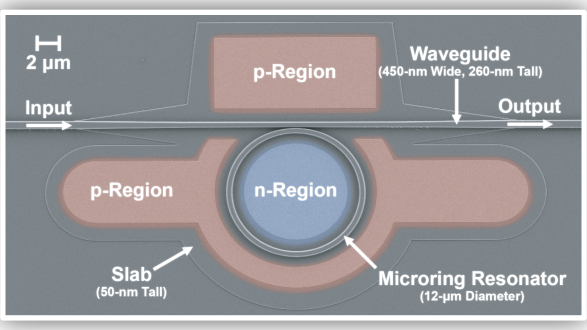
18 GHz



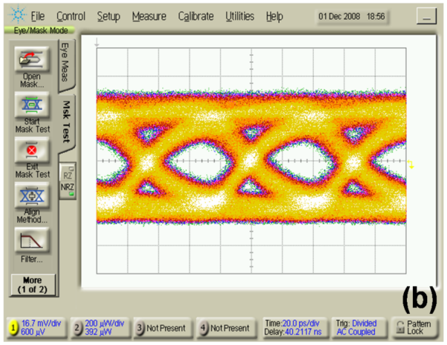


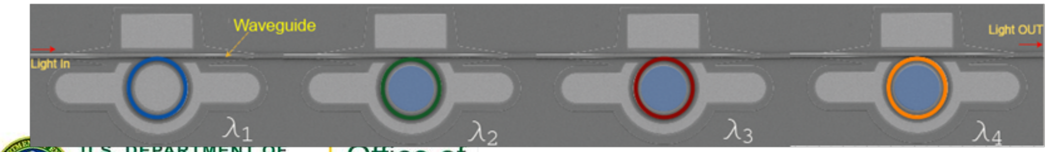


Multi-Wavelength Receiver Array



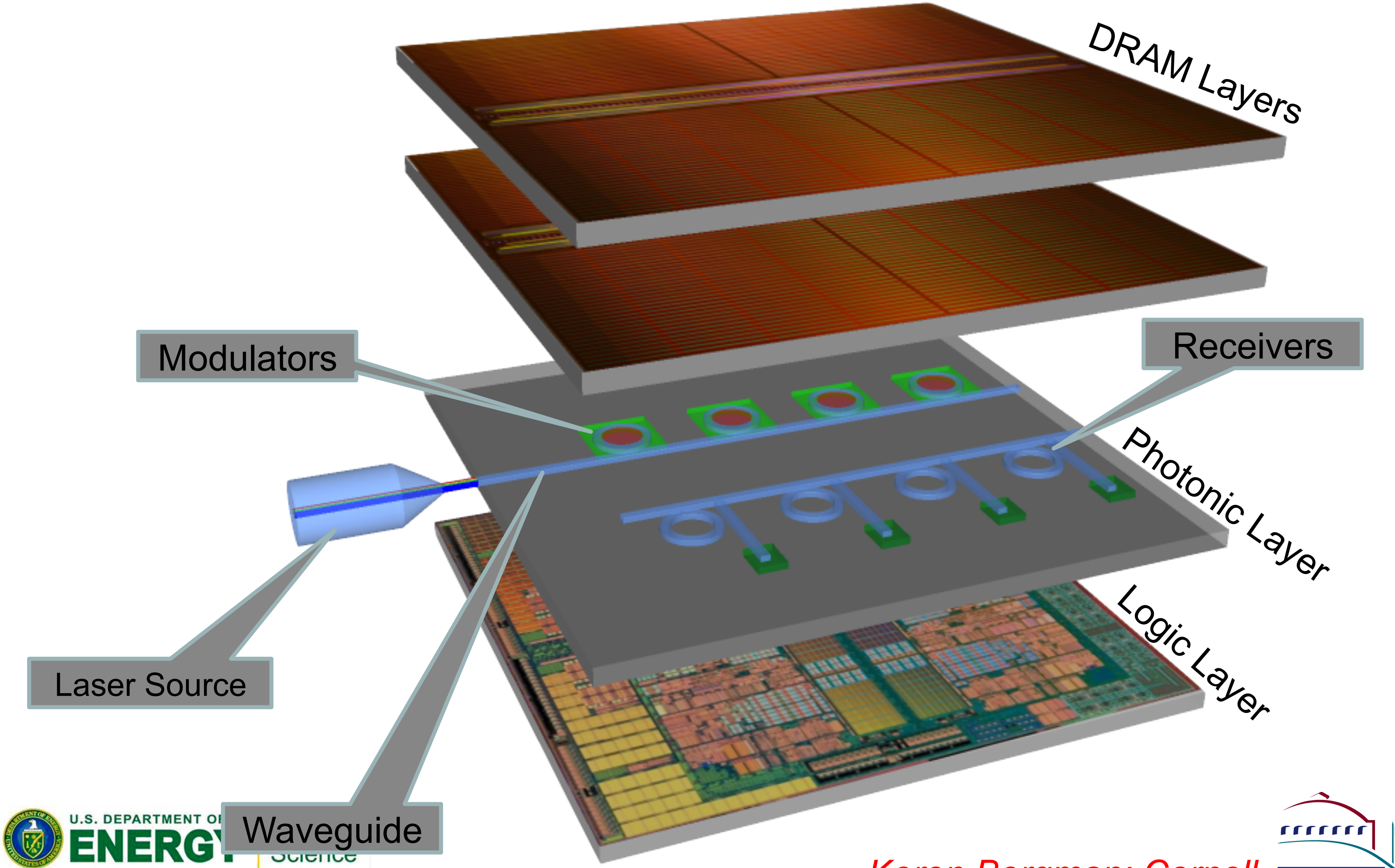
Multi-Wavelength Modulator Array





Demux data wavelength channel λ_4 at 15 Gbps

Stacked Logic with Integrated Silicon Photonics



Keren Bergman: Cornell

Silicon Photonics: Optical Lambda Switching integrated on CMOS Chips

- Silicon Photonics enables WDM optical switching “Fabric” integrated directly with CMOS logic (grand unification)
  - Lambda switching in solid-state (no MEMS or diffraction gratings)
  - Optics finally moving “on-chip” to break through pin-limits
- Similar to current WAN scale lambda switching
  - Grand-unification of on-CMOS-chip and off-chip optical switching to minimizes OEO conversions
  - Need protocol for managing virtual circuits and packet routing tables together (GMPLS)
  - QoS management is similar to OSCARS service (but on-chip)
- If we actually have dedicated end-to-end lambdas, why use AIMD protocol to manage the flow rate?
  - Particularly between resources within a datacenter
  - Infrastructure for fixed-datarate protocols (with OSCARS)
  - Unification with flow-control and QoS mgmt on HPC system

System on Chip (SoC) integration Moving the NIC on Chip

- **Moore's Law continues** *(but what should we do with those transistors?)*
  - Could use it to cram more cores on chip, Or more cache
  - Or integrate other components (SoC) such as NIC
  - PCIe is wasted in cloud where nodes connected to ethernet fabric +disk in most cases (move features on chip to reduce cost)
- **Cloud and Consumer market drivers for SoC Integration**
  - Already see PCIe and 10GigE has moved on chip in commodity space (10G on BG/P, Niagara, and latest Intel Sandybridge. 100GigE by 2018??)
  - Vendors will ask you "which NIC" should we put on board?
    - cloud is pushing for ethernet (standards based interconnect)
  - *At high-end the "custom interconnect" is the "converged fabric" (e.g. Power7) with re-provisioning of pins for PCIe/Ethernet*
- ***What would you do with 100Gig NIC on each chip?***
  - *Coordinated data transfers from each node?*
  - *Is the "network the computer" or the "computer is the network?"*

Exascale I/O

I/O Technology
(HEC-FSIO Discussion)

- **Mechanical Disk storage: spindle limited**
  - Requires exponentially more devices (more subject to failure)
  - Need to purchase more capacity than we want to get bandwidth
- **NVRAM/FLASH: way faster than disk, but expensive**
  - Can easily purchase sufficient bandwidth
  - But cannot afford the capacity that we need
- ***Gary Grider's "Reese's Peanut Butter Cup" solution: Hybrid I/O with NVRAM for defensive I/O that bleeds off to disk***
- **Shared Filesystems vs. Distributed Filesystems**
  - Difficult to scale POSIX consistency model to exascale
  - Consider how to integrate node-localized storage into hierarchy
  - How does one manage a distributed filesystem?

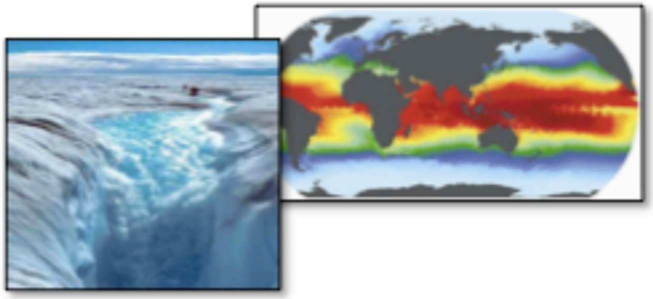
Other I/O Issues

- **Defensive I/O (for ~10x higher MTTI)**
  - **Localized Checkpointing:** SCR to local NVRAM could supply required bandwidth
  - *How does one manage node-distributed persistent storage?*
- **Analysis I/O**
  - **In-situ (locality aware) data analysis**: e.g. MapReduce: Layout data across cluster and ship computation to the storage (functional semantics)
  - **Object database storage** (HDF, NetCDF) pushed into the storage infrastructure (interoperate with locality-aware storage)
- **Data provenance**
  - As we move to analysis of experimental data, need to know who touched the data and when (NASA example)
  - Requires coordination with data transport infrastructure

Application Drivers

DOE mission imperatives require simulation and analysis for policy and decision making

- *Climate Change*: Understanding, mitigating and adapting to the effects of global warming
  - Sea level rise
  - Severe weather
  - Regional climate change
  - Geologic carbon sequestration
- *Energy*: Reducing U.S. reliance on foreign energy sources and reducing the carbon footprint of energy production
  - Reducing time and cost of reactor design and deployment
  - Improving the efficiency of combustion energy sources
- *National Nuclear Security*: Maintaining a safe, secure and reliable nuclear stockpile
  - Stockpile certification
  - Predictive scientific challenges
  - Real-time evaluation of urban nuclear detonation



Accomplishing these missions requires exascale resources.

Uncertainty Quantification for Predictive Simulation

- Want to go from an ability to describe natural phenomena with simulations towards a *predictive capability*
  - But nature is messy: need to understand sensitivity to preturbation
  - Numerical simulation answers whether a design is sufficient, but does not quantify the uncertainty of the answer.
  - This is NOT V&V *(can only do UQ if you trust your simulation)*
  - Example Application: *rapid qualification of new nuclear power plant design, or many engineering problems*
- Example Approach: *Polynomial Chaos*
  - Run many simulations with input preturbations *(task sched/mgmt)*
  - Statistical summarization across simulation datasets to understand sensitivity to design parameters *(huge data management issues)*
- Requires workflow tools integrated with transport infrastructure
  - Need task farming to prevent batch system from being overwhelmed (need task management & data management)
  - Need coordination with network infrastructure, I/O, and compute
  - *No pretty graphical tools (get over that now!)*

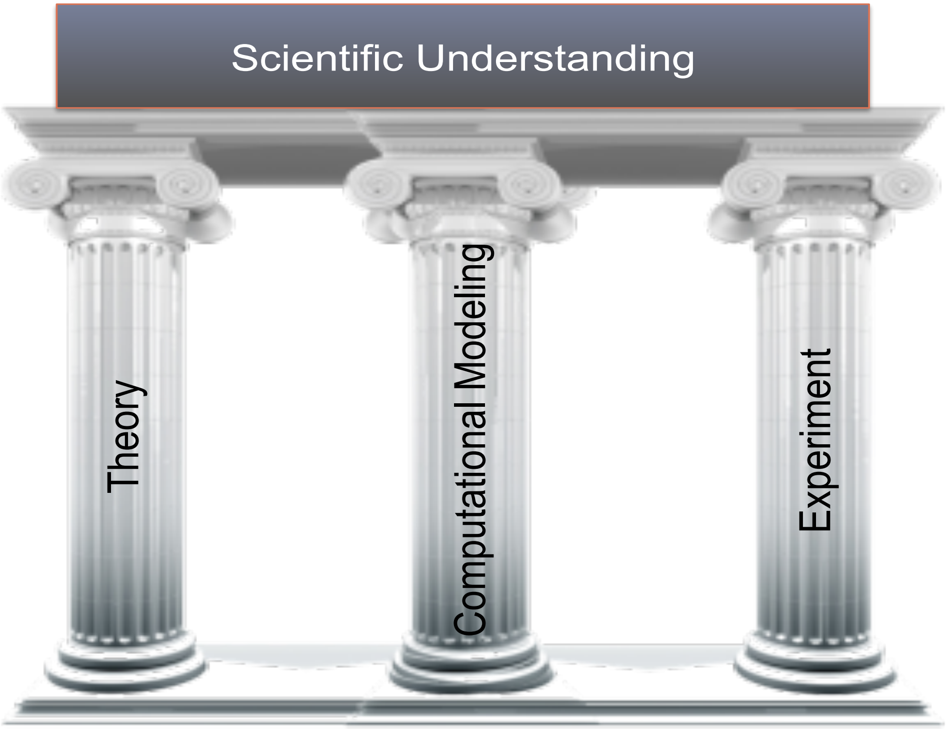
The 3 Pillars of Science

(High End Computing Revitalization Task Force, D. Reed, 2003)

- ***Theory*: mathematical models of nature**

- ***Experiment*: empirical data about nature**

- ***Computation*: enables mathematical models to be applied to complex phenomena that are closer to experiment & nature.**



- **Predictive modeling requires tight integration of these 3 pillars!**
  - **Computational models are used to test theories involving complex phenomena that cannot be matched directly against experiments**
  - **Enable comprehension of complex experimental data**

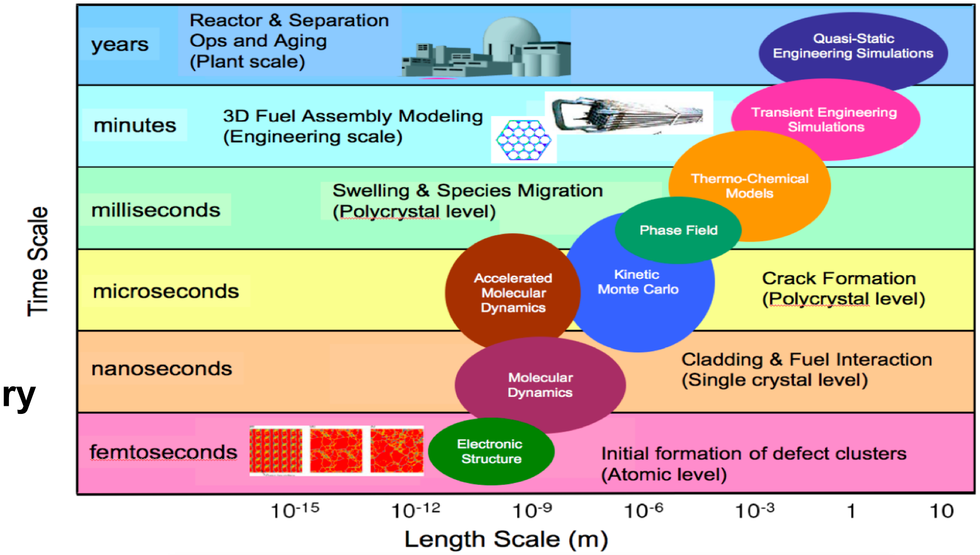
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Predictive simulations are a critical capability for nuclear energy

(Koonin 2010)

- **Key science and engineering challenges**
  - Life-time extension of light water reactor
    - 3d fuel failure
    - Evolution of pin and assembly failure
  - Modular reactor design and new fuels
    - Fluid/structure interactions
    - Full scale plant radiation field modeling
- **Reducing uncertainty through improved theory and simulation**
  - Cross-section methods, variance and usage
  - Up-scaling micro to macro structures
  - 3d thermomechanics and swelling
  - Fission gas release and migration at microscale
  - Atomistic-to-3D macroscale simulation
- **Impact**
  - 20% reduction in cost of each nuclear plant
  - Increase operating margins to increase safety
  - Reduce uncertainty for existing reactors
  - Enable insertion of new fuel technology in existing reactors
  - Speed licensing of new designs



Time Scale

years: Reactor & Separation Ops and Aging (Plant scale); Quasi-Static Engineering Simulations

minutes: 3D Fuel Assembly Modeling (Engineering scale); Transient Engineering Simulations

milliseconds: Swelling & Species Migration (Polycrystal level); Thermo-Chemical Models; Phase Field

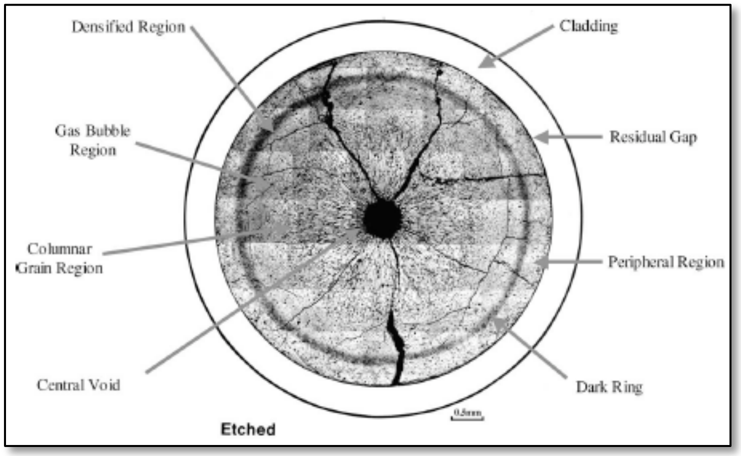
microseconds: Accelerated Molecular Dynamics; Kinetic Monte Carlo; Crack Formation (Polycrystal level)

nanoseconds: Molecular Dynamics; Cladding & Fuel Interaction (Single crystal level)

femtoseconds: Electronic Structure; Initial formation of defect clusters (Atomic level)

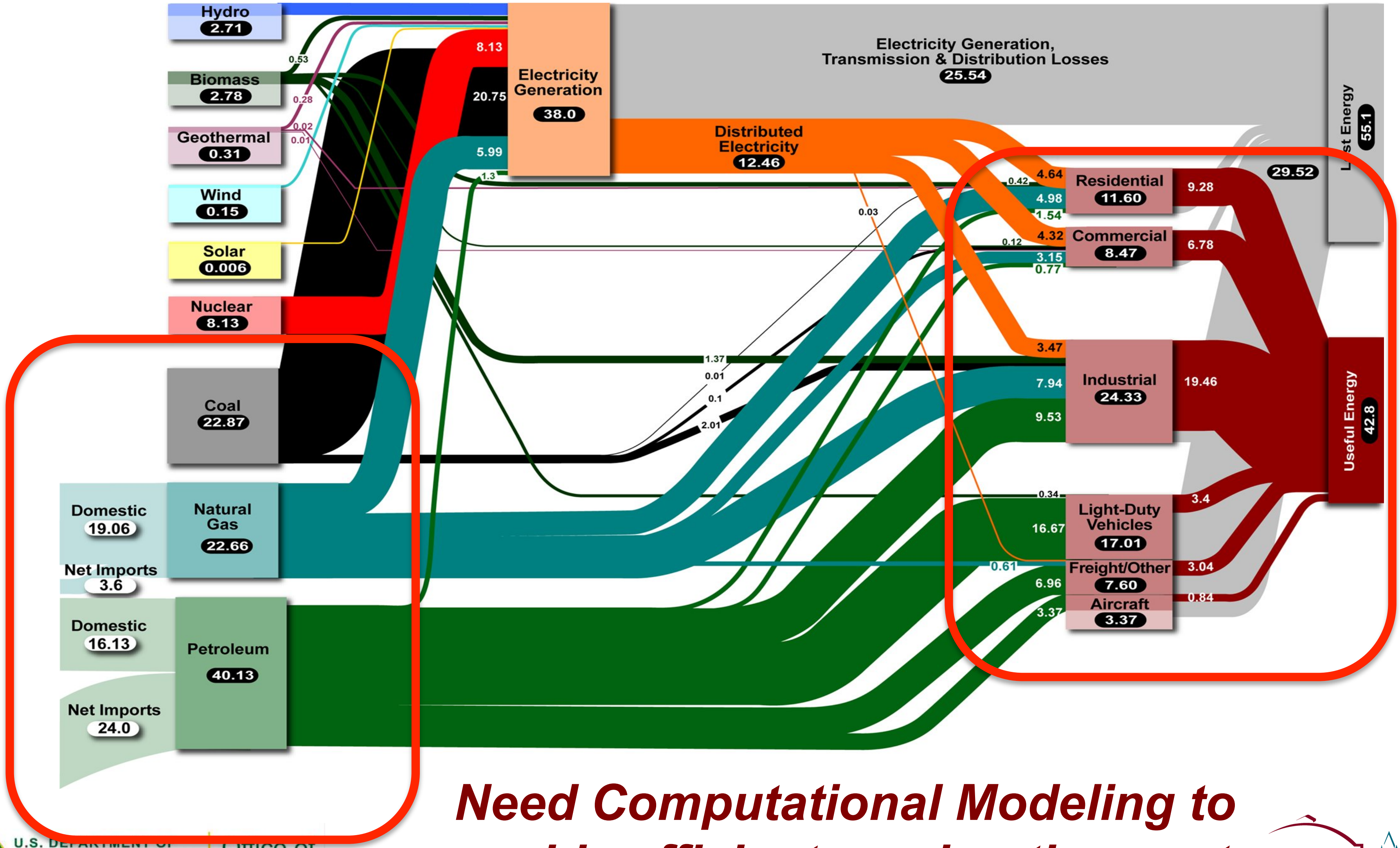
10^{-15} 10^{-12} 10^{-9} 10^{-6} 10^{-3} 1 10

Length Scale (m)



Fuel microstructure: from Wolf, BES-SciDAC workshop

Combustion accounts for 85% of the energy used in the United States.



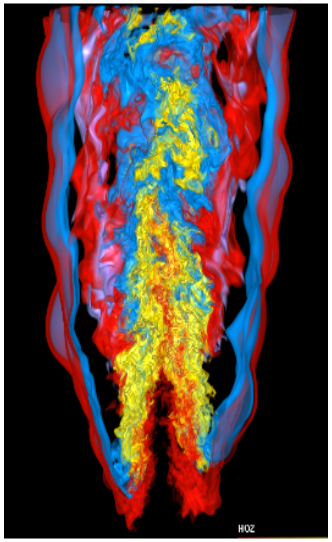
Need Computational Modeling to enable efficient combustion systems

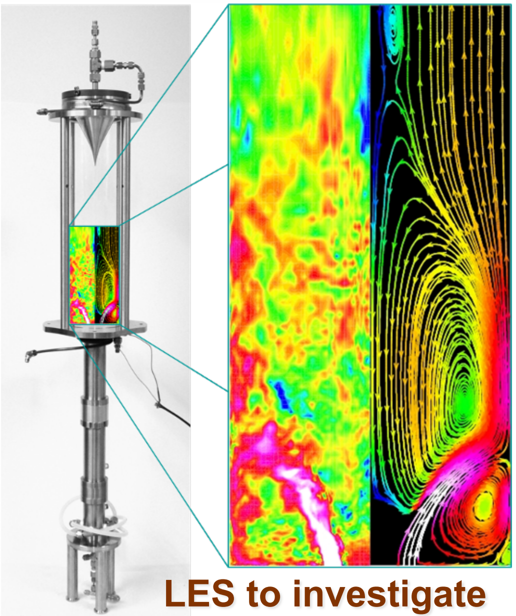
Soucrce: LLNL

High End Modeling and Data Assimilation For Advanced Combustion Research

Approach: **Combine unique codes and resources to maximize benefits of high performance computing for turbulent combustion research**

Advanced "capability-class" solvers





DNS to investigate combustion phenomena at smallest scales

no modeling

limited applicability

LES to investigate coupling over full range of scales in experiments

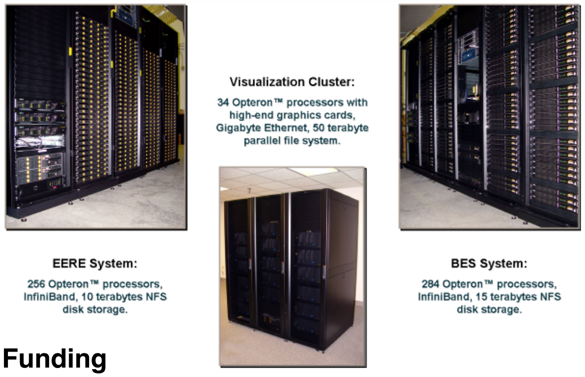
minimal modeling

full geometries

Access to leading edge computational resources

CRF Computational Combustion and Chemistry Laboratory

Combustion Research and Computational Visualization Facility



Visualization Cluster:

34 Opteron™ processors with high-end graphics cards, Gigabyte Ethernet, 50 terabyte parallel file system.

EERE System:

256 Opteron™ processors, InfiniBand, 10 terabytes NFS disk storage.

BES System:

284 Opteron™ processors, InfiniBand, 15 terabytes NFS disk storage.

Joint OS-EERE Funding

DOE Office of Science Laboratories

LBNL NERSC

ORNL OLCF

ANL ALCF

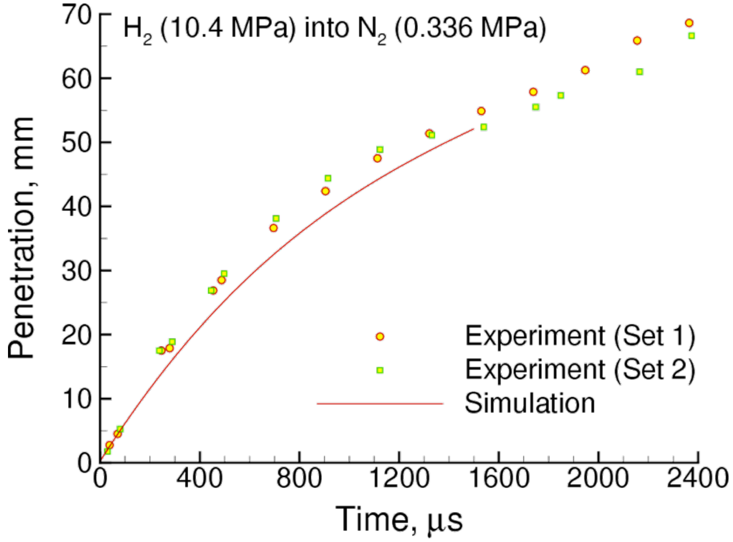
INCITE Program

U.S. DEPARTMENT OF ENERGY | Office of Science

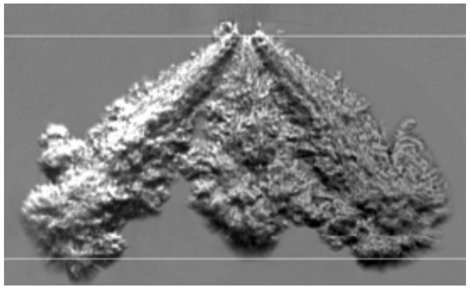
Ofelein, Chen: Sandia 2009

Example of HPC for Predictive Modeling

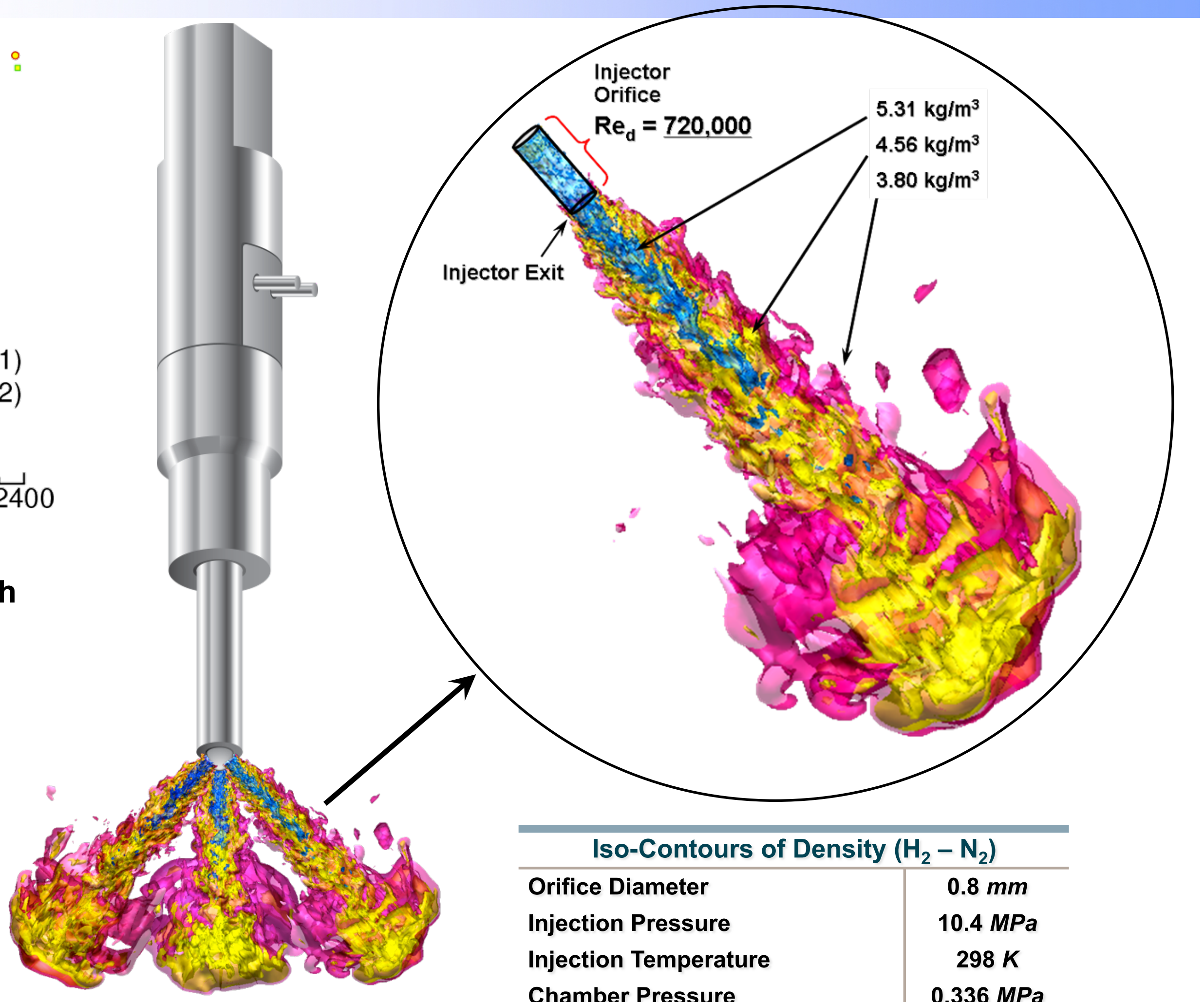
(*Rigorous validation of high-pressure injection)*



Representative comparison of LES with penetration measurements



Shadowgraph (U. Wisconsin)



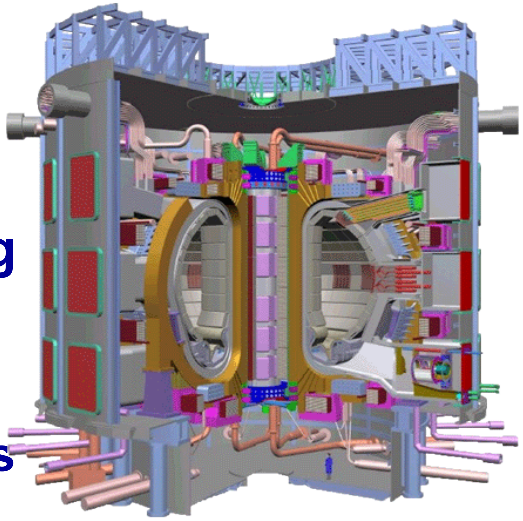
Large Eddy Simulation

Iso-Contours of Density (H_2 – N_2)	
Orifice Diameter	0.8 *mm*
Injection Pressure	10.4 *MPa*
Injection Temperature	298 *K*
Chamber Pressure	0.336 *MPa*
Chamber Temperature	298 *K*

Ofelein: Sandia 2009

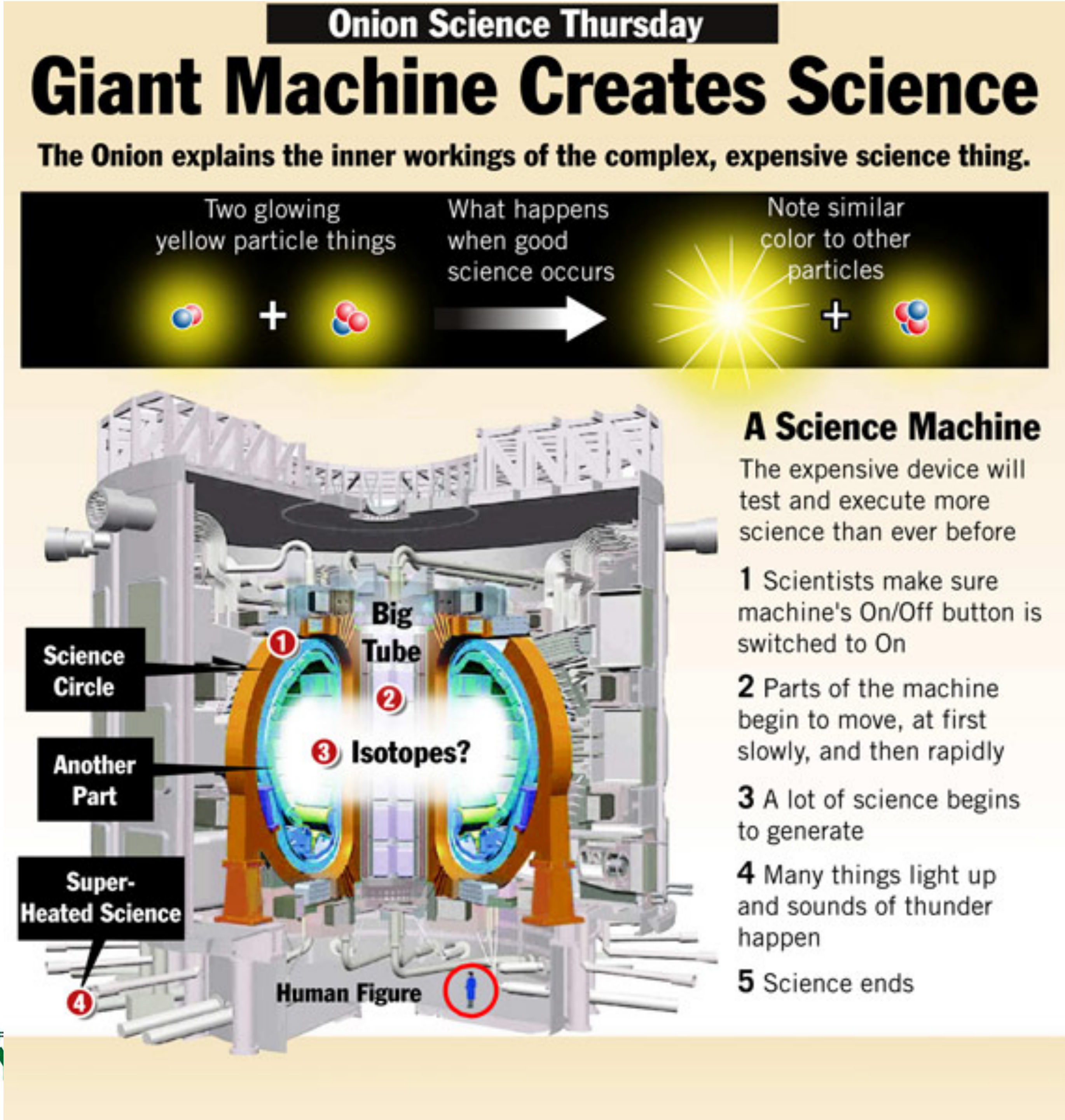
Fusion

Towards Whole Device Modeling Capability



ITER: International Thermonuclear Experimental Reactor

- **Fusion science has been dominated by scaling first-principles models of specific phenomena**
  - **Dozens of independent codes focused on narrow area**
- **ITER development requires full-device modeling capability by 2018**
  - **For shot planning and device control**
  - **Requires Code-coupling, Multi-scale multiphysics**
  - **Uncontrolled discharge could damage $12B device!**
- **Requires new code and algorithms to span 12 orders magnitude in time and length scales (Keys/Jardin)**
  - Exaflop-scale hardware capability as a minimum requirement (3 orders of magnitude)
  - Requires complementary Math/CS investments in algorithms and software infrastructure (9 orders of magnitude)



Onion Science Thursday

Giant Machine Creates Science

The Onion explains the inner workings of the complex, expensive science thing.

Two glowing yellow particle things
+
What happens when good science occurs
Note similar color to other particles
+
Science Circle
Another Part
Super-Heated Science
Big Tube
1
2
3 Isotopes?
4
Human Figure

A Science Machine

The expensive device will test and execute more science than ever before

1 Scientists make sure machine's On/Off button is switched to On

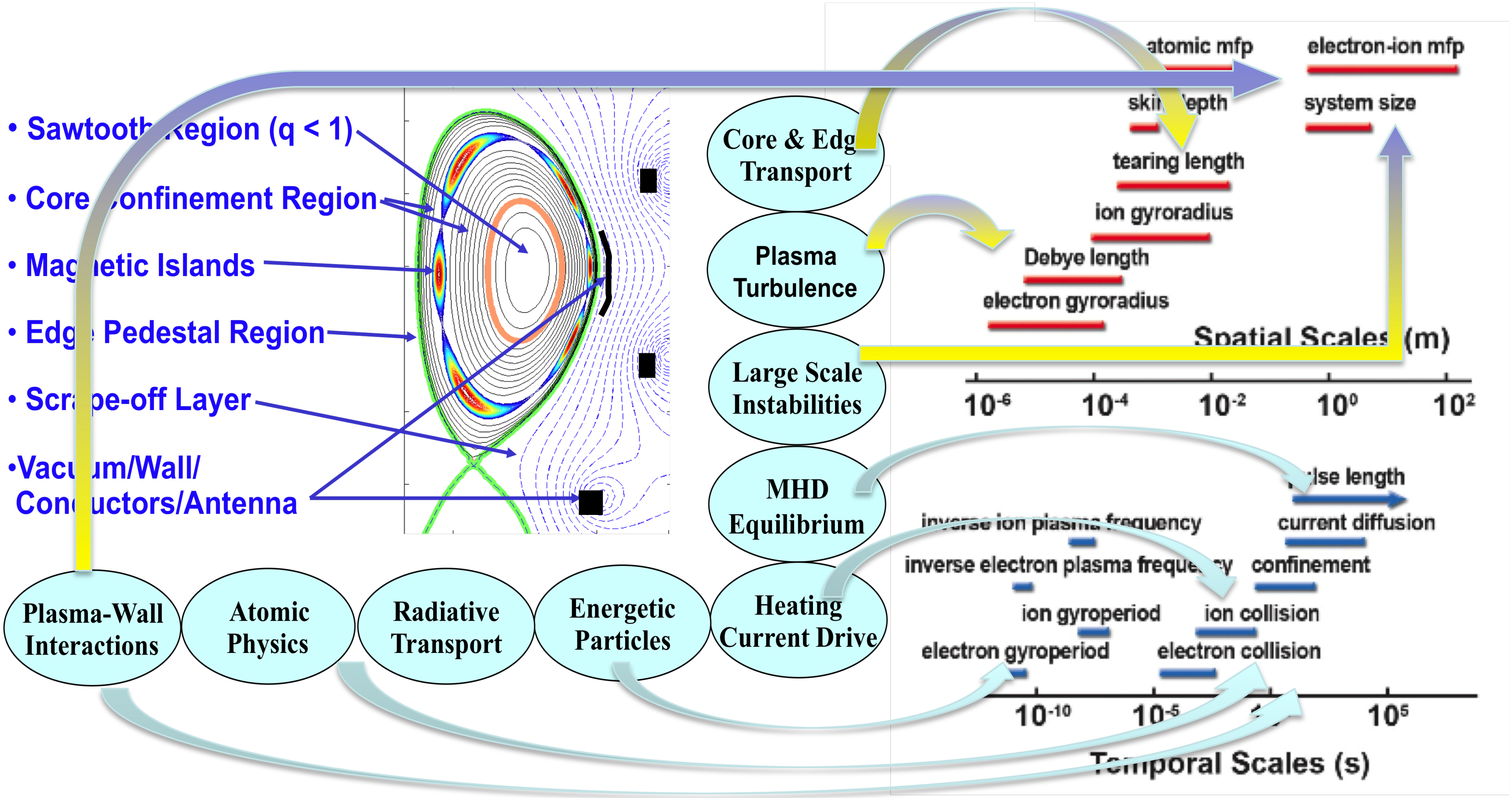
2 Parts of the machine begin to move, at first slowly, and then rapidly

3 A lot of science begins to generate

4 Many things light up and sounds of thunder happen

5 Science ends

Full Device Modeling: Complex Multiphysics Interactions

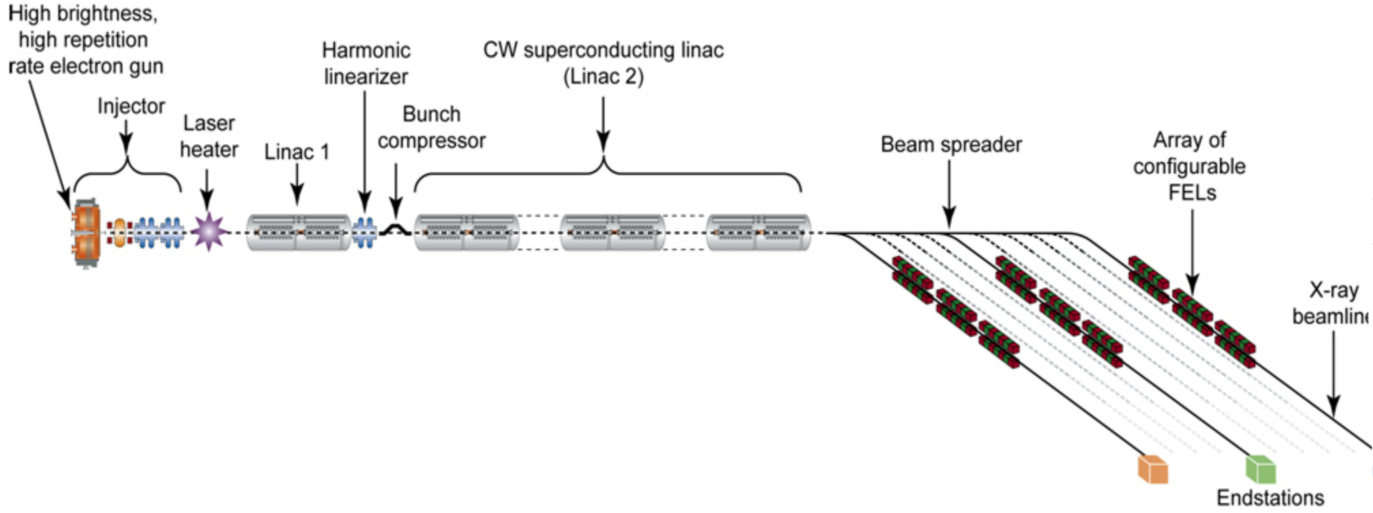


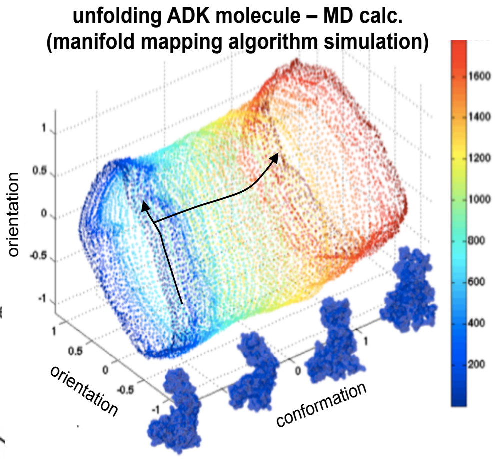
- Complex multiphysics interactions between key components of Tokamak requires models that span 12 orders of magnitude (time and length scales)

Risk to Program if Predictive Simulation Capability is Not Available

- Uncontrolled discharge
  - ITER good for 200 experiments (less if loss of plasma confinement)
  - Can destroy $12B device in a single uncontrolled event
  - Predictive modeling for shot-planning is critical to prevent such events
- US Participation in ITER project
  - Access to ITER experiment will be gated by ability to plan useful experiments
  - US access requires US leadership in simulation capability
- DEMO engineering design/planning
  - Next fusion device after ITER for sustained magnetically confined fusion
  - Understanding data collected from ITER experiments requires analytical modeling capability
  - Predictive modeling and simulation is essential component for controlling engineering costs and risk

Next Generation Light Source: Tomographic Image Reconstruction


High brightness, high repetition rate electron gun
Injector
Laser heater
Linac 1
Harmonic linearizer
Bunch compressor
CW superconducting linac (Linac 2)
Beam spreader
Array of configurable FELs
X-ray beamline
Endstations


unfolding ADK molecule – MD calc.
(manifold mapping algorithm simulation)

- **Computational requirements JUST for orientation reconstruction**
  - ***Input Data Rate:*** **10^5 images/second at 10^6 pixels imaging rate (4TB/sec)**
  - **10^5 of images of diffraction patterns representing 2D projection of the sample in random orientation**
  - **Best available orientation algorithms require ~N^6 flops (N=1000 for NGLS detector)**
  - ***Total performance required is 10^{18} FLOP/s for pulse rate of 10^5 images/second***
- **Similar requirements for shot planning**

Both data processing and shot planning will require exascale computing for analysis and terabit networking for data movement

Data Intensive Computing, Shot Planning, and Data Re-Analysis

- **Know that data rates from experiments are increasing at a dramatic rate**
  - **WW-LHC Computing Grid, PLANCK are existing examples with primarily 1-way information flow for data analysis**
  - **New examples of massive data sources with ITER, JGI, and NGLS emerging with massive flows both ways for data assimilation and shot planning, and re-analysis**

- **Turn-around for experiments limited by**
  - **Data movement rate (networking resources)**
  - **Throughput for data analysis**
  - **Throughput to run simulations to plan next shot**
  - ***Ability to process data and plan experiments will limit access to the device***

Overall Conclusions

- **Future of computing is power limited**
  - Limited by end of Dennard scaling for logic
  - Limited by energy cost of moving bits
  - Result is 1000x increase in parallelism and constrained bandwidth
  - Massive changes open up many new opportunties

- **Technology Opportunities**
  - System on Chip Integration: Every chip might have an ethernet NIC on-board (is the network the computer or is the computer the network?)
  - Silicon Photonics (grand unification of optics with CMOS, solid state lambda switching with no OEO conversions, massive all-optical lambda-switching fabric)

- **Application Opportunities**
  - Coupled multi-component multiphysics applications
  - Uncertainty Quantification and Predictive Modeling
  - Increased need to compare theory to experiment (massive data flows)
  - Increased need for bi-directional interactions with experiments for “shot planning” (analyze and then simulate with fast turn-around)

More Info

- DOE Exascale Workshops Series
  - http://extremecomputing.labworks.org/
- International Exascale Software Project (IESP)
  - http://www.exascale.org/

Bonus Material

Exascale Architecture Constraints

System attributes	2010	"2015"		"2018"	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200 GB/sec	
MTTI	days	O(1day)		O(1 day)	

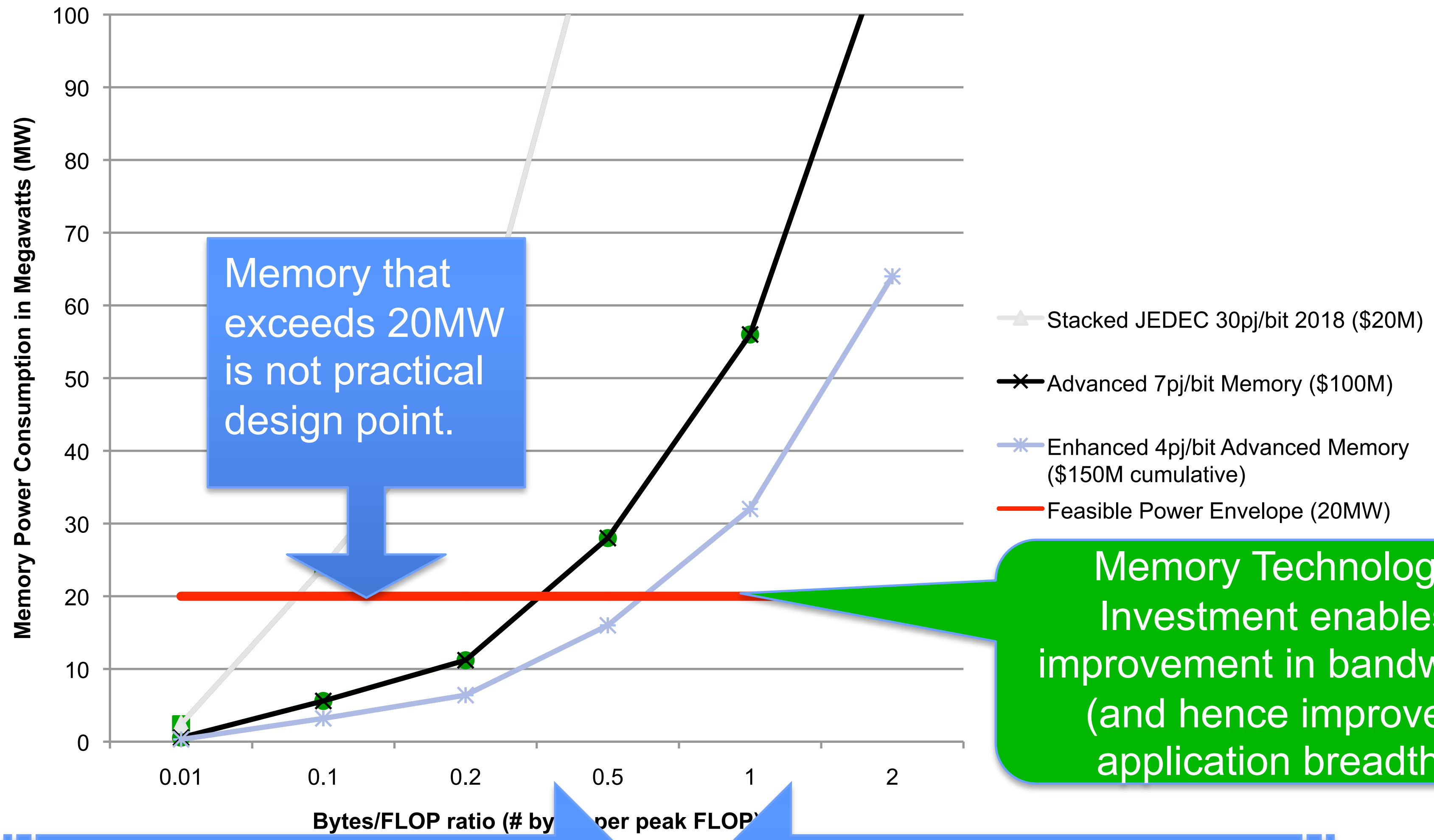
Exascale Initiative Steering Committee
(circa December 9, 2009)

Systems	2009	2015 +1/-0	2018 +1/-0
System peak	**2 Peta**	**100-300 Peta**	**1 Exa**
Power	**6 MW**	**~15 MW**	**~20 MW**
System memory	0.3 PB	5 PB	64 PB (+)
Node performance	125 GF	0.5 TF or 7 TF	1-2 or 10TF
Node memory BW	25 GB/s	1-2TB/s	2-4TB/s
Node concur	12	O(100)	O(1k) or 10k
Total Node Intercon	3.5	[illegible]	00GB/s 4 or 1:8 from memory
System size (no	[illegible]	[illegible]	00) or O(1M)
Total concurren	[illegible]	[illegible]	O) for latency hiding
Storage	15 P	150 P	00 PB (>10x system memory is min)
IO	0.2 TB	10 TB/s	60 TB/s (how long to drain the machine)
MTTI	days	O(1day)	O(1 day)

60 MW over budget

OOOPs!

Limiting Memory Bandwidth Limits System Scope



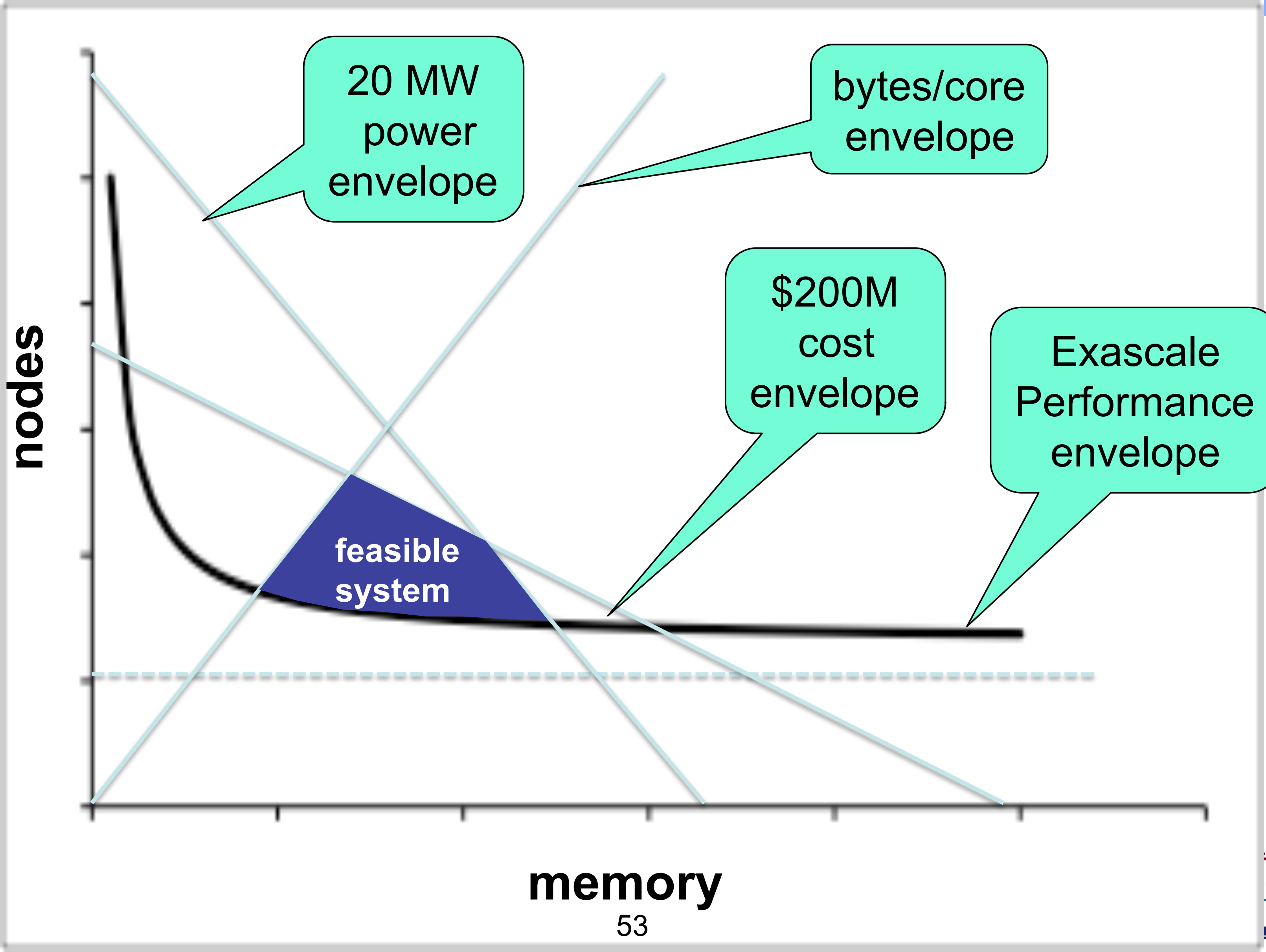
Memory that exceeds 20MW is not practical design point.

Memory Technology Investment enables improvement in bandwidth (and hence improves application breadth)

Application performance and breadth pushes us to higher BW

Power pushes us to lower bandwidth

Using Co-Design to Navigate a Complex Trade-Space



Interesting Architecture Trends that Might Intersect with Terabit Networking

- 2018-2020 may be the transition point of seeing optics move on-chip

- Moore’s Law continues
  - Could use it to cram more cores on chip
  - Or more cache
  - Or perhaps improve integration of other components (SoC) such as NIC

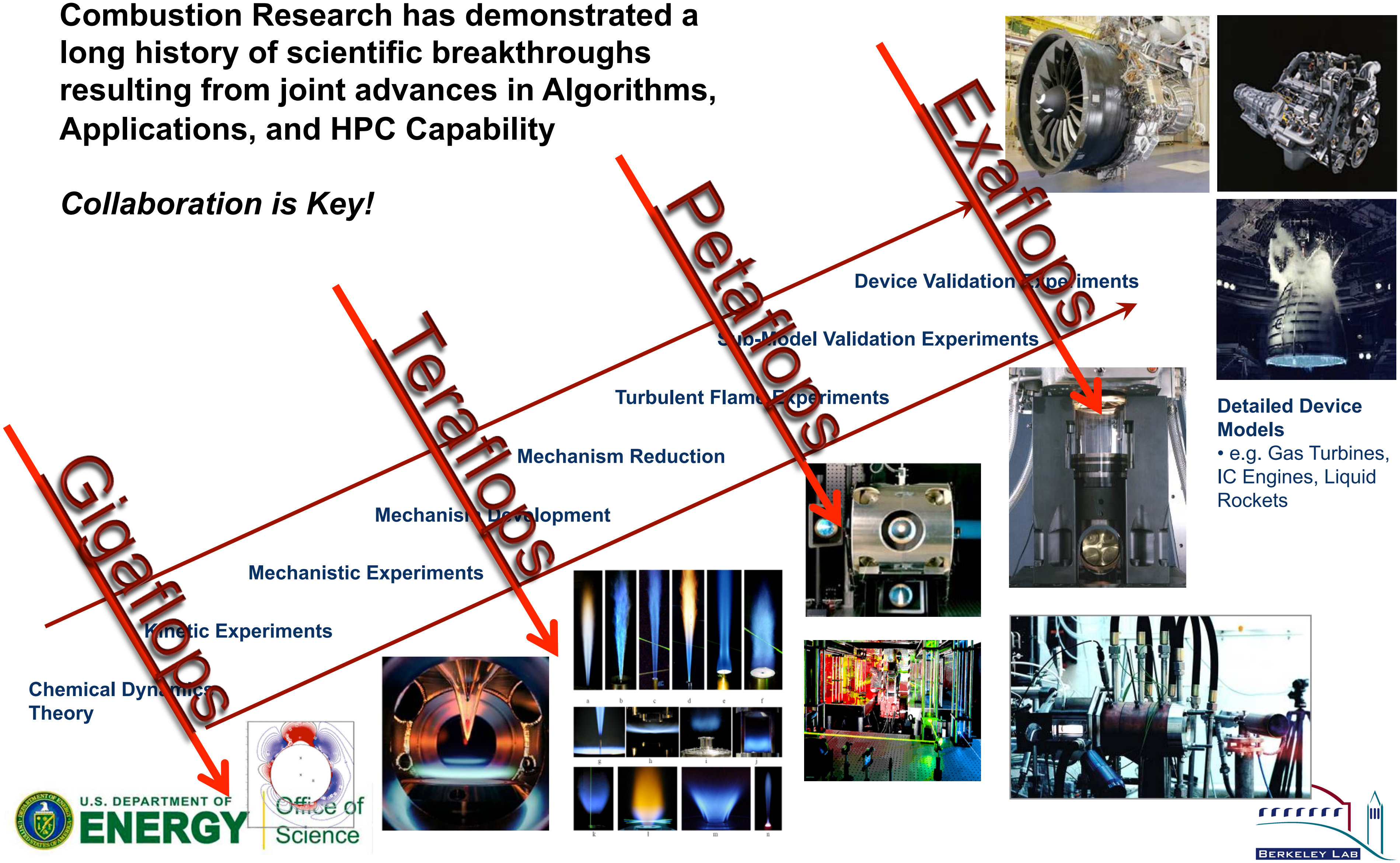
- What can you do with optics on chip?

- What can you do if very node has a 100Gigabit NIC on board every single socket in the system?

Scientific Breakthroughs Enabled by Algorithms, Applications, and HPC Capability

Combustion Research has demonstrated a long history of scientific breakthroughs resulting from joint advances in Algorithms, Applications, and HPC Capability

Collaboration is Key!



Gigaflops

Teraflops

Petaflops

Exaflops

Chemical Dynamics Theory

Kinetic Experiments

Mechanistic Experiments

Mechanism Development

Mechanism Reduction

Turbulent Flame Experiments

Sub-Model Validation Experiments

Device Validation Experiments

Detailed Device Models
- e.g. Gas Turbines, IC Engines, Liquid Rockets

JGI/Bioinformatics

- Database Access
  - Need Guaranteed QoS for big query responses (not traditional download and analyze)
- Re-Analysis
  - Searching for matches against current database of sequences (using BLAST)
  - Periodic “sanity checking” of currently stored data
- Data Provenance
  - Need to know who inserted the data and when
  - Constant annotation of stored data

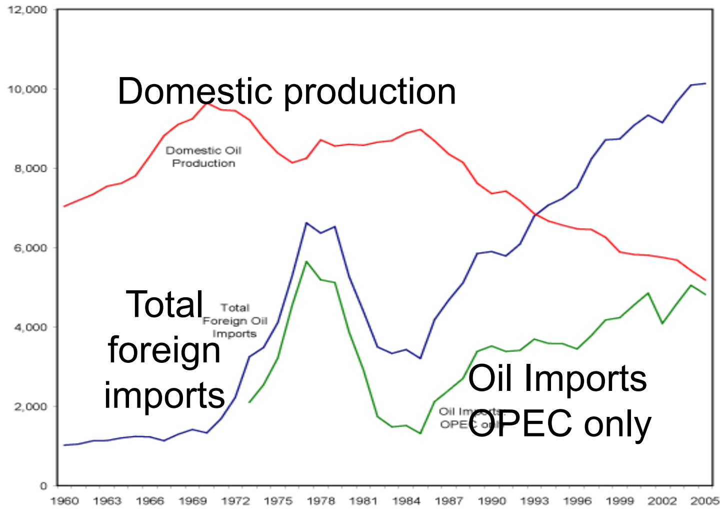
DOE Mission Drivers for Extreme Scale Computing

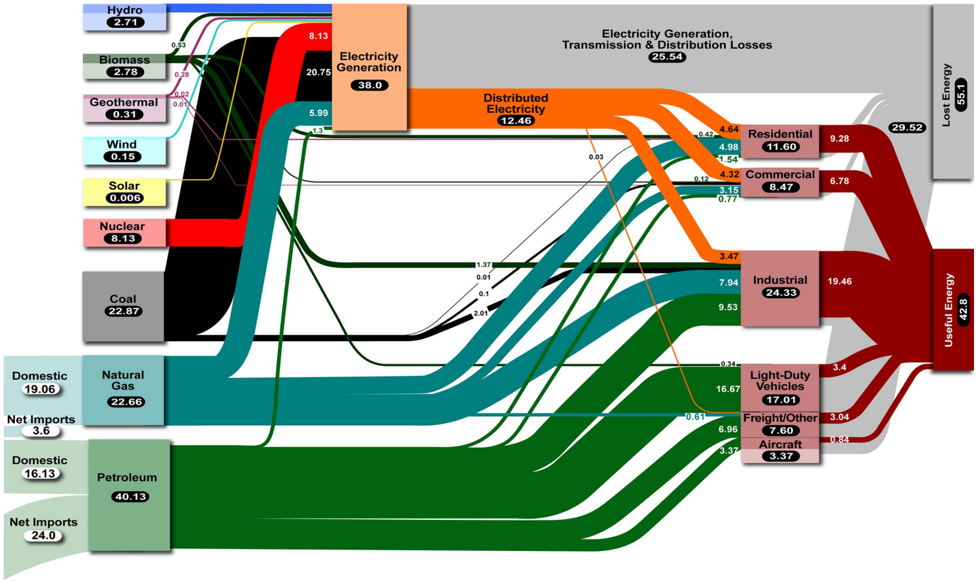
- National Security
  - dependence on unreliable sources
- Economic Security
  - need for assured supplies at affordable prices
- Environmental Security
  - obtaining energy in ways that does not harm the environment

Koonin, ASCAC 2009

US Oil Production and Foreign Oil Imports

(thousands of barrels per day)





Data Intensive Computing for Exascale Applications

- Predictive Simulation and Uncertainty Quantification
  - Engineering Simulation for rapid qualification of new nuclear reactor designs or design optimization
  - Workflows and integration
- Multiphysics Simulations
  - However, “heterogeneous computing” may not be as heterogeneous as you might think
- Data Analysis for large experiments
  - PPDG, Climate, JGI and PLANCK are current examples
- Shot planning for large experiments
  - Make the most of very expensive experimental apparatus
  - ITER, Light Sources